Atelier GAST 2024

Gestion et Analyse de données Spatiales et Temporelles

Organisateurs :

Nida MEDDOURI (LRE, EPITA, Kremlin-Bicêtre, France)
Loïc SALMON (ISEA, Université de Nouvelle-Calédonie, France)
Aurélie LEBORGNE (ICube, Université de Strasbourg, France)

PRÉFACE

Le neuvième atelier « Gestion et Analyse des données Spatiales et Temporelles » (GAST) est associé à EGC'2024. Cet atelier, s'appuyant sur le groupe de travail GAST, regroupe des chercheurs, du domaine académique et de l'industrie, qui s'intéressent aux problématiques liées à la prise en compte de l'information temporelle ou spatiale (quantitative ou qualitative) dans leurs processus de gestion et d'analyse de données (méthodes et application d'extraction, de gestion, de représentation, d'analyse et de visualisation d'informations).

Ces actes regroupent huit soumissions présentées à l'atelier GAST'2024. Ces articles montrent une large étendue des recherches actuelles à des fins de modélisation, d'extraction, d'analyse, ou de visualisation d'information, basées sur les dimensions temporelles et spatiales associées. Nous espérons que les orateurs, les auditeurs et les lecteurs pourront interagir autour de ces sujets, que les questions et les défis associés à l'information temporelle et spatiale continueront à animer les débats.

Nous tenons à remercier tous les auteurs pour leurs propositions d'articles ainsi que les membres du comité de lecture dont les retours ont été de qualité pour l'ensemble des articles.

En espérant que ces articles vous apporteront de nouvelles perspectives autour de la Gestion et l'Analyse des données Spatiales et Temporelles, nous vous souhaitons une bonne lecture.

Nida Meddouri — LRE, EPITA
Loïc Salmon — ISEA, UNC
Aurélie Leborgne — ICube, UNISTRA

Membres du comité de lecture

Achraf Mtibaa (MIRACL, Université de Sfax, Tunisie)
Albrecht Zimmermann (GREYC, Université de Caen Normandie, France)
Aurelie Leborgne (ICube, Université de Strasbourg, France)
Baptiste Lafabregue (ICube, Université de Strasbourg, France)
Clément Iphar (LETG, Université de Bretagne Occidentale, France)
Cyril de Runz (LIFAT, Université de Tours, France)
Florence Le Ber (ICube, Université de Strasbourg/ENGEES, France)
François Rioult (GREYC, Université de Caen Normandie, France)
Jérôme Gensel (LIG, Université Grenoble Alpe, France)
Jimmy Randrianasoa (LRE, EPITA, France)
Loïc Salmon (ISEA, Université de Nouvelle-Calédonie, France)
Ludovic Moncla (LIRIS, INSA Lyon, France)
Nazha Selmaoui-Folcher (ISEA, Université de Nouvelle-Calédonie, France)
Nida Meddouri (LRE, EPITA, France)
Pedro Merino Laso (École Nationale Supérieure Maritime, France)
Roberto Interdonato (CIRAD, UMR TETIS, France)
Thomas Guyet (INRIA, AIstroSight, France)
Thomas Lampert (ICube, Université de Strasbourg, France)

TABLE DES MATIÈRES

Analyse de données de piraterie en mer pour l'observation des zones maritimes à risque

Loïc Salmon*, Pedro Merino Laso**

*Université de la Nouvelle-Calédonie, - BP R4 98851 Nouméa Cedex, Nouvelle-Calédonie
ISEA (Institut des Sciences Exactes et Appliquées)
loic.salmon@unc.nc
**French Maritime Academy (ENSM), 1 Rue de la Noë, 44300 Nantes, France
IRENav, EA 3634, BRCM Brest, Ecole Navale C600, 29240 Brest cedex 9, France
pedro.merino-laso@supmaritime.fr
https://www.supmaritime.fr/

La piraterie maritime est un phénomène ancien qui a perduré à travers les âges évoluant avec le temps pour s'adapter aux nouvelles réalités géopolitiques et technologiques. Ce fléau est caractérisé par des actes de violence, de vol et de détournement de navires en haute mer. Au cours des dernières décennies, la piraterie maritime a attiré l'attention mondiale notamment en raison des attaques répétées au large des côtes somaliennes ciblant essentiellement des navires commerciaux pour rançons, créant ainsi une crise maritime internationale et amenant des travaux scientifiques sur ces données de piraterie Marchione et Johnson (2013). De plus, les progrès technologiques ont également modifié la nature de la piraterie, avec l'utilisation de moyens plus sophistiqués tels que la cyberpiraterie, qui vise à compromettre la sécurité des systèmes de navigation et de communication des navires. Cette forme de piraterie met en évidence la nécessité d'adapter les réglementations et les technologies de sécurité pour faire face aux nouvelles menaces émergentes Merino Laso et al. (2021). Dans le cadre de ce travail, une analyse des données de piraterie en mer est faite pour observer l'évolution spatio-temporelle ainsi que la nature des attaques (armes utilisées, type de navire victime de l'attaque) afin de déterminer les zones à risque ainsi que la nature de ce risque.

1 Données de piraterie en mer

Les données utilisées dans cette étude proviennent de plusieurs sources et ont été croisées de manière similaire aux travaux de Li et Yang (2023) utilisant trois jeux de données différents.

Le premier correspond aux données collectées auprès du Bureau maritime international (IMB), nettoyées et enrichies de données géospatiales. Ce jeu de données contient des informations sur les attaques de pirates maritimes survenues entre janvier 1993 et décembre 2020, ainsi que des données d'indicateurs de pays pour la même période Benden et al. (2021).

Le deuxième jeu de données Piracy and Armed Robbery (PAR) associé au Global Integrated Shipping Information System (GISIS) ont été collectées par l'Organisation Maritime Internationale (OMI), et sont disponibles sur `https://gisis.imo.org/Public/PAR/Default.aspx`. L'ensemble de données PAR couvre les cas d'attaques de pirates survenues

dans le monde entier du 1er janvier 2006 à aujourd'hui. Les incidents de piraterie sont encodés avec de nombreuses caractéristiques, notamment l'emplacement, la date, les détails de l'incident et les conséquences de l'incident.

Enfin, le troisième jeu de données correspond aux données de piraterie de l'ASAM (Anti-Shipping Activity Messages) disponibles à l'adresse suivante `https://msi.nga.mil/Piracy` et donne des informations concernant le type de navire visé, la date, la position, ainsi que la description de l'attaque.

L'ensemble de ces données sont croisées et fusionnées dans la suite, évidemment avec des informations incomplètes pour certaines attaques et des données de qualité variable de manière générale.

2 Analyse et traitement des données de piraterie

Pour faire de la prédiction concernant les zones à risque et la nature du risque encouru par le navire en fonction de sa taille ou de sa situation de navigation notamment, nous envisagions de considérer l'ensemble des données. Seulement, en observant les zones de piraterie, celles-ci évoluent au cours du temps, et certaines données anciennes ne sont plus pertinentes. Une analyse spatio-temporelle de ces données est donc nécessaire afin d'observer l'évolution des zones de piraterie ainsi que des moyens mis en place par les pirates afin de déterminer les données pertinentes pour faire de la prédiction. Des travaux précédents concernent l'analyse spatio-temporelle de données de piraterie en mer Li et Yang (2023), seulement l'évolution des zones de piraterie n'y est pas considérée, les auteurs analysant d'une part la composante spatiale (zones à risque et clusters sur l'ensemble des données indépendamment de la date) et d'autre part la composante temporelle (saisonnalité des attaques, heures des attaques) sans évaluer ou constater l'évolution de ces zones au cours du temps.

3 Observation de l'évolution des zones de piraterie

Une illustration de l'évolution de ces zones est représentée dans la Figure 1. Celle-ci montre en vert les positions où ont eu lieu des attaques entre 2004 et 2009 (non inclus) et en orange celles ayant eu lieu entre 2009 et 2014 (non inclus). On constate au niveau du Golfe de Guinée, que les attaques entre 2004 et 2009 se restreignaient uniquement aux côtes, tandis qu'entre 2009 et 2014 les attaques se sont éloignées des cotes. De la même façon, on observe que les attaques entre 2004 et 2009 avaient lieu en majorité dans le Golfe d'Aden contre partout dans la zone entre 2009 et 2014. On observe donc une évolution des zones d'acte de piraterie qui dans le cas du Golfe de Guinée s'est déportée pour s'éloigner de la côte, et dans le cas du Golfe de l'Aden s'est étendu très au-delà.

Évidemment, des processus de fouille de données doivent être mis en place pour constater et détecter ces évolutions, à partir de méthodes dérivées de DBSCAN Birant et Kut (2007) ou Kernel Density Backurs et al. (2019). L'évolution des actes de piraterie en mer dépend de l'augmentation des moyens que les pirates ont à disposition pour attaquer leur cible, mais également du comportement des navires qui prennent leur précaution quand ils arrivent dans des zones à risque ou bien les évitent. De plus, d'autres champs sont à considérer parmi lesquels le type de navire attaqué (bateau de passager, tanker ...), le type d'attaque des pirates, la situation

de navigation du navire (au mouillage, au bord des côtes, en pleine mer), ainsi que la description complète de l'attaque (mentionnant entre autres suivant la qualité des données, le nombre d'assaillants, leur embarcation, etc...). Il s'agira d'étudier une partie de ces problématiques, à savoir d'évaluer l'évolution des zones maritimes à risque et d'observer l'influence du type de navire et du statut de navigation et voir si elles sont en lien avec les aspects spatio-temporels.

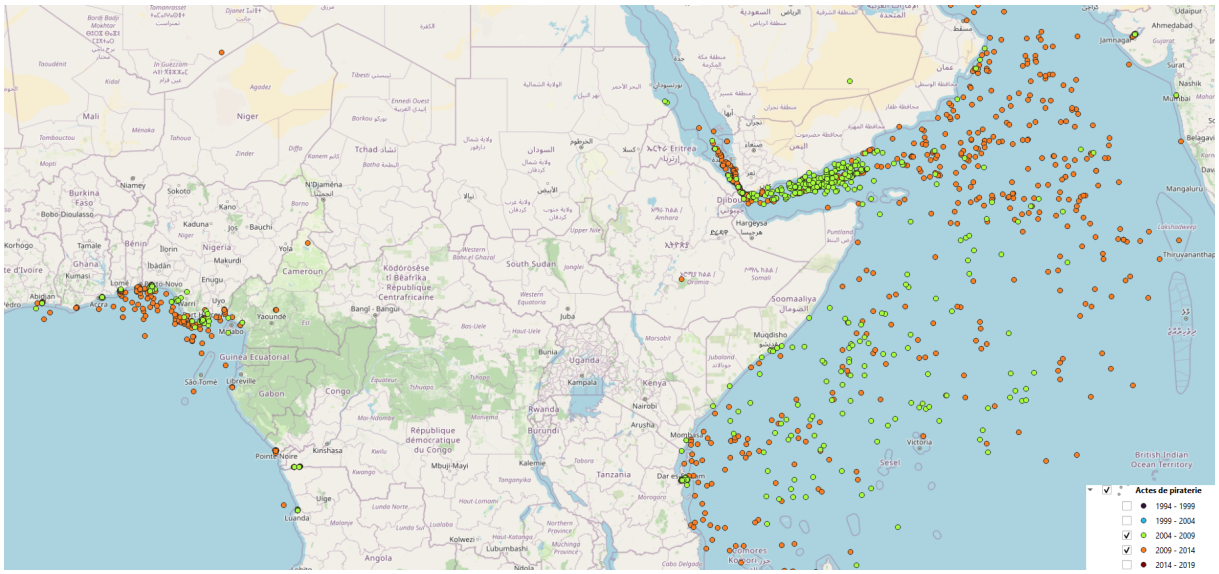


FIG. 1 – *Exemple d'évolution des zones de piraterie entre les périodes 2004-2009 (vert) et 2009-2014 (orange)*

Références

Backurs, A., P. Indyk, et T. Wagner (2019). Space and time efficient kernel density estimation in high dimensions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, et R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Benden, P., A. Feng, C. Howell, et G. V. Dalla Riva (2021). Crime at sea : A global database of maritime pirate attacks (1993–2020). *Journal of Open Humanities Data*.

Birant, D. et A. Kut (2007). St-dbscan : An algorithm for clustering spatial–temporal data. *Data Knowledge Engineering 60*(1), 208–221. Intelligent Data Mining.

Li, H. et Z. Yang (2023). Towards safe navigation environment : The imminent role of spatio-temporal pattern mining in maritime piracy incidents analysis. *Reliability Engineering System Safety 238*, 109422.

Marchione, E. et S. D. Johnson (2013). Spatial, temporal and spatio-temporal patterns of maritime piracy. *Journal of Research in Crime and Delinquency 50*(4), 504–524. PMID : 25076796.

Merino Laso, P., L. Salmon, M. Bozhilova, I. Ivanov, N. T. Stoianov, G. Velev, C. Claramunt, et Y. Yanakiev (2021). ISOLA : An innovative approach to cyber threat detection in cruise shipping. *Smart Innovation, Systems and Technologies*.

Summary

L'analyse et la fouille de données est un élément important pour comprendre l'évolution des phénomènes dans l'espace et le temps. Une analyse spatio-temporelle des actes de piraterie en mer est nécessaire afin de voir l'évolution de ces actes au cours du temps si bien par leur nature que les zones sur lesquelles ils se produisent. Dans cet article, nous proposons des éléments pour traiter ces données à caractère spatio-temporel afin de déterminer les zones à risque.

ShiftDTW : adaptation de la métrique DTW pour la clusterisation de séries temporelles cycliques

Lucas Foulon*, Ilyes Korichi*
Xavier Millot**

*Data-Major
prenom.nom@data-major.com,
https://www.data-major.com/
**Oxtaam
prenom.nom@oxtaam.com
https://www.oxtaam.com/

Résumé. L'élasticité de la métrique DTW permet une comparaison plus souple entre les séries temporelles, et est utilisée dans de nombreux domaine de l'apprentissage automatique, comme la classification ou la clusterisation. Cependant, elle ne permet pas de faire coïncider les mesures de début et de fin de séries temporelles si elles possèdent un décalage qui intervient dès le début d'une des séries et dont la partie omise se retrouve à la fin de cette dernière. Dû à la cyclicité de ce type de séries — qui ne possèdent ni début ni fin — nous nous appuyons sur les travaux Cyclic DTW pour proposer une approximation moins coûteuse de cette méthode de calcul qui sera ensuite utilisée avec la méthode de clustering K-Means.

1 Introduction

Les séries temporelles sont présentes dans de nombreux contextes tel que les cours des actions boursières, les données météorologiques, ou encore les mesures biomédicales (Aghabozorgi et al., 2015). L'enjeu du clustering est de trouver des similitudes entre ces données tout en catégorisant différemment les données les plus éloignées. L'objectif de notre projet consiste à regrouper au mieux les entreprises ayant des comportements comptables similaires. Ces comportements sont basés sur les indicateurs comptables de chaque entreprise pour trouver les meilleurs regroupements possibles. Les indicateurs comptables représentent des montants financiers, des données numériques qui évoluent dans le temps et qui permettent de mesurer la santé financière d'une entreprise. Cependant, toutes les entreprises n'ont pas la même ancienneté, et pour les comparer, nous normalisons chaque indicateur sous la forme d'une année "type", représentant au mieux le comportement de l'indicateur sur une année. Cette étape de normalisation peut être faite de différentes façons : dans notre cas, nous utilisons la représentation de la saisonnalité fournie par l'outil FACEBOOK PROPHET (Taylor et Letham, 2018) permettant ainsi d'uniformiser la durée de tous les indicateurs de toutes les entreprises. En effet, PROPHET a été développé pour obtenir des prévisions sur les séries temporelles ayant de

fortes saisonnalités (quotidienne, hebdomadaire, annuelle, etc). À partir du modèle généré, il est possible d'en extraire la tendance et les saisonnalités associées.

Mais quelle que soit la méthode d'uniformisation utilisée, nous souhaitons mettre en évidence des similitudes qui existeraient à des saisons différentes. Par exemple, une entreprise de sport d'hiver et une entreprise de sport d'été peuvent avoir des comportements comptables similaires mais décalés de 6 mois. De la même façon, certains producteurs de fruits et légumes réalisent certaines transactions selon des saisons bien précises : bien que leurs comportements similaires ne se déroulent pas sur les mêmes saisons, nous souhaiterions alors pouvoir regrouper ces producteurs. Contrairement à la méthode *Cyclic Dynamic Time Warping* (CDTW) proposé par Palazón-González et Marzal (2012), nous ne souhaitons pas garder une élasticité totale : nous utilisons une bande de Sakoe-Chiba permettant de réduire les comparaisons extrêmes — celles qui provoqueraient l'omission de nombreuses valeurs consécutives durant la comparaison entre les deux séries — et nous n'attendons pas à connaître avec précision le meilleur alignement, mais le meilleur à quelques pas de temps près, au profit d'un allègement du temps de calcul. En effet, les méthodes comme K-Means demandent de recalculer les distances vers les nouveaux barycentres à chaque itération, et cela peut devenir assez coûteux.

Nous proposons une méthode ayant la même complexité de calcul que DTW, qui ne retournera pas l'alignement optimal, mais qui testera plusieurs alignements possibles. Ces alignement sont strictement définis par la taille de la bande Sakoe-Chiba utilisée. C'est notamment grâce à cette bande que nous limitons la complexité de notre méthode, comme présenté dans la partie 4.

2 État de l'art

En plus de la distance euclidienne, il existe de nombreuses mesures de distance adaptées aux séries temporelles. En effet, leur nature continue nous pousse à analyser leurs similarités de façons différentes que d'autres types de données numériques. Par exemple, nous pouvons citer la distance Kullback-Leibler (Warren Liao, 2005), la mesure de la *Longest Common Subsequence* ou bien encore la méthode *Dynamic Time Warping* (DTW) (Sakoe et Chiba, 1971). Cette dernière permet de donner de l'élasticité à l'axe du temps, que l'on pourrait sinon trouver trop rigide dans certains cas. Elle a été rendue populaire pour la reconnaissance de la parole avec Sakoe et Chiba (1971, 1978) puis de nouveau comme métrique de similarité pour les séries temporelles (Keogh et Pazzani, 2000).

Plusieurs variantes de DTW ont été proposé depuis, pour améliorer les performances de calcul (Salvador et Chan, 2007) ou convenir à une mesure de similarité plus adapté au contexte des séries temporelles. Sakoe et Chiba (1978) proposent d'utiliser une bande appelée Sakoe-Chiba pour limiter la taille de la déformation de DTW, et donc de réduire le coût de calcul DTW. On retrouve aussi le *Derivative Dynamic Time Warping* (Keogh et Pazzani, 2001) n'utilise plus la distance euclidienne pour la comparaison pair à pair, mais la différence des dérivées estimées en tout point pour chacune des séries. Dans notre contexte, les séries temporelles que nous manipulons n'ont ni début ni fin, car bien que temporelles, chacune de ces séries représente la saisonnalité annuelle d'un indicateur numérique. La méthode "naïve" consisterait à mesurer plusieurs fois les distances entre les deux séries en décalant une des deux séries pour chaque pas de temps, mais cela augmenterait considérablement le coût de calcul. Pour la dis-

tance d'édition [1] (ED), Maes (1990) démontre qu'il est possible de réduire le coût en utilisant une approche "divide-and-conquer". Palazón-González et Marzal (2012) proposent la mesure CDTW, s'inspirant de l'algorithme de Maes (1990), avec un coût de $\mathcal{O}(mn \times log(m))$ avec m et n la taille des deux séquences. Avec les solutions trouvées sur les séquences de départ, il est possible de borner la matrice des distances dans les itérations suivantes. Une fois bornée, les mêmes séquences sont comparées mais avec l'une d'elles décalée dans le temps. La méthode retournera l'alignement des deux séquences qui obtient le plus petit score DTW. Comme Maes (1990), les auteurs prouvent formellement que leur méthode retourne le même résultat que la méthode "naïve", mais avec un coût moindre. Dans notre contexte d'application, nous n'avons pas besoin de cette garantie, et nous privilégions la réduction du coût de calcul à la précision du résultat.

Nous allons définir les concepts essentiels dans la partie 3 pour pouvoir présenter notre mesure de distance dans la partie 4. Enfin nous verrons les résultats obtenus dans la partie 5 avant de conclure (partie 6).

3 Définitions

Série temporelle. Une séquence ou une série temporelle T de longueur n est définie par une suite de valeurs $(t_0, t_1, \ldots, t_{n-1})$ avec $t_i \in \mathbb{R}$.

Dynamic Time Warping. Le calcul de la distance DTW entre deux séries T et S, respectivement de longueur m et n, se base sur la matrice des distances $M_{(T,S)}$ entre ces 2 séries tel que $M_{(T,S)}[i,j] = |t_i - s_j|$ avec $0 \leq i < m$ et $0 \leq j < n$. La complexité de calcul est $\mathcal{O}(n * m)$. L'alignement optimal est défini par le chemin de déformation le plus court depuis la matrice des distances cumulées, calculée à partir de $M_{(T,S)}$ tel que :

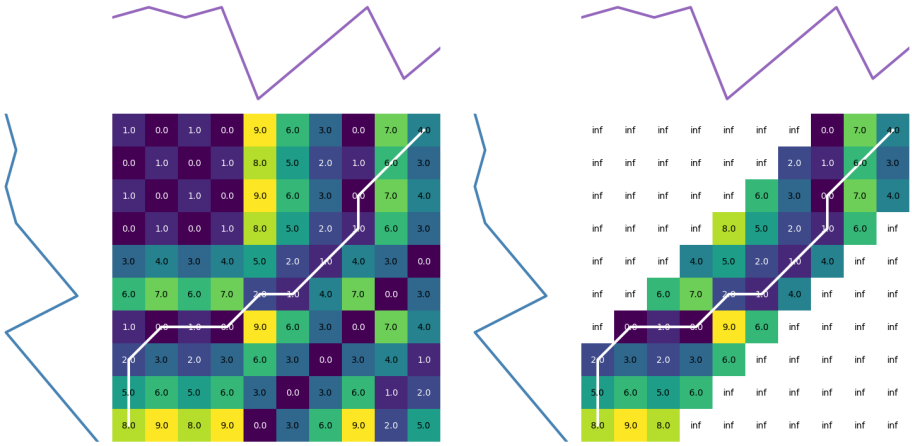
$$Mcumul_{(T,S)}[i,j] = \begin{cases} M_{(T,S)}[0,0] & \text{si } i = j = 0, \\ M_{(T,S)}[i,j] + min \begin{cases} Mcumul_{(T,S)}[i-1,j] \\ Mcumul_{(T,S)}[i,j-1] \\ Mcumul_{(T,S)}[i-1,j-1] \end{cases} & \text{sinon.} \end{cases} \tag{1}$$

La figure 1a présente la matrice de distance entre deux séries temporelles, ainsi que le chemin le moins coûteux déterminé par DTW. La distance euclidienne emprunterai simplement la diagonale de cette matrice pour déterminer son résultat. Avec un masque de Sakoe-Chiba, la complexité de calcul serait réduite car nous ne parcourons plus l'ensemble de la matrice des distances, comme présenté dans la figure 1b.

Série temporelle cyclique. Une séquence ou une série temporelle T de longueur n peut subir un changement cyclique σ tel que $\sigma(t_0, t_1, \ldots, t_{n-1}) = (t_1, \ldots, t_{n-1}, t_0)$.

Comme défini par Palazón-González et Marzal (2012), notons σ^k l'arrangement d'une série temporelle de k décalages, alors deux séries temporelles sont équivalentes s'il existe une

1. appelée aussi "distance de Levenshtein"



(a) Matrice des distances entre deux séries temporelles.(b) Matrice des distances entre deux séries temporelles avec un masque Sakoe-Chiba.

FIG. 1 – Matrices des distances entre deux séries temporelles. Les deux séries possèdent les mêmes valeurs, mais l'une d'elle a été décalé dans le temps.

valeur de $k \in \mathbb{N}$ tel que $T = \sigma^k(T')$. Alors la série temporelle cyclique T est notée :

$$[T] = \{\sigma^k(T) : 0 \leq k < m\} \tag{2}$$

Cyclic Dynamic Time Warping. La mesure CDTW (Palazón-González et Marzal, 2012) entre $[T]$ et $[S]$ est défini tel que :

$$CDTW([T],[S]) = \min_{0 \leq k < n} \left(\min_{0 \leq l < n} DTW(\sigma^k(T), \sigma^l(S)) \right) \tag{3}$$

Dans leurs travaux, les auteurs démontrent qu'il est possible de réduire l'équation pour obtenir :

$$CDTW([T],[S]) = \min_{0 \leq k < n} \left(\min(DTW(\sigma^k(T), S), DTW(\sigma^k(T)t_k, S)) \right) \tag{4}$$

avec $\sigma^k(T)t_k$ la concaténation de l'élément t_k avec $\sigma^k(T)$.

Avec cette simplification, il propose un algorithme de type "*branch-and-bound*" qui permet de réduire la complexité à $\mathcal{O}(mn \times log(m))$ avec m et n la taille des deux séquences. Ils précisent également les ce coût ne peut être réduit grâce à l'utilisation de bande de Sakoe-Chiba. Dans notre cas, nous souhaitons réduire le coût au maximum et il ne nous est pas nécessaire de tester toutes les solutions incluses dans $CDTW([T],[S])$. Nous proposons dans la partie suivante de limiter les solutions avec une bande de Sakoe-Chiba.

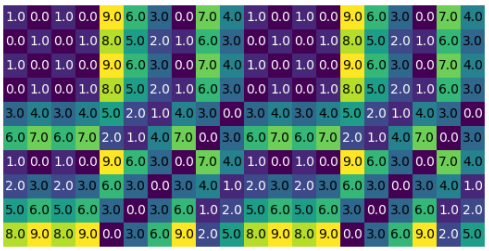


FIG. 2 – La matrice des distances doublée entre deux séries temporelles.

(a) Sans décalage, première itération.

(b) Avec un décalage, seconde itération.

FIG. 3 – Matrices des distances entre deux séries temporelles avec un masque de Sakoe-Chiba. La matrice de la première itération est similaire au calcul DTW avec masque présenté dans la figure 1b. La seconde itération retourne un score plus petit et donc meilleur.

4 Présentation de la mesure ShiftDTW

Dans cette partie, nous allons décrire les étapes de la mesure ShiftDTW et la modification apportée à K-Means pour prendre en considération cette nouvelle mesure.

Notons T et S de longueur m et n, deux séries temporelles à comparer. L'algorithme 1 décrit l'ensemble des étapes nécessaires au calcul ShiftDTW entre T et S avec une bande Sakoe-Chiba de taille r. Dans un premier temps, nous calculons les distances entre les valeurs pair-à-pair des deux séries temporelles (la matrice calculée est un des paramètres de l'algorithme 1). Nous obtenons une matrice des distances de tailles $m \times n$ que nous doublons afin d'obtenir une matrice de taille $2 \times m \times n$. Avec l'exemple de la figure 1a, nous obtenons le résultat présenté dans la figure 2.

Ensuite, comme pour DTW, la matrice des distances cumulées sera calculée comme décrit dans l'équation 1, en tenant compte de la bande Sakoe-Chiba. Cette première itération est en tout point similaire à DTW classique avec la même bande Sakoe-Chiba (figure 3a). Les itérations suivantes sont effectuées sur la même matrice de distances cumulées mais en partant avec un décalage égal à celui de la taille de la bande Sakoe-Chiba (figure 3b).

Pour chacune des itérations, un chemin — le plus court — est calculé, et la mesure ShiftDTW retourne le plus court d'entre eux avec la taille du décalage associé.

Algorithme 1 : Calcul de la mesure ShiftDTW entre les séries T et S sachant r.

Data : $M_{(T,S)}$ la matrice des distances pair-à-pair entre T et S
Data : r la taille du rayon de la bande Sakoe-Chiba
Data : l_T la longueur de la série temporelle T
Résultat : La mesure $ShiftDTW_r$ entre T et S, et la valeur de décalage d
$ShiftDTW_r \leftarrow \infty$;
$d \leftarrow None$;
$M_{(2T,S)} \leftarrow concat(M_{(T,S)}, M_{(T,S)})$; `/* taille` $m \times n \rightarrow 2 \times m \times n$ `*/`
$Mask \leftarrow calcul\_masque\_Sakoe\_Chiba(r)$;
pour $0 \leq i < l_T$ *par pas de* $2 \times r + 1$ **faire**
  $Mcumul_{(T,S)}[:] = cumul\_depuis\_matrix\_dist(M_{(2T,S)}[i : l_T + i], Mask)$;
  `/* calcul matrice des distances cumulées */`
  $current\_dist = \sqrt{Mcumul_{(T,S)}[-1, -1]}$;
  if $current\_dist < ShiftDTW_r$ **then**
    $ShiftDTW_r \leftarrow current\_dist$;
    $d \leftarrow r$;
  end
fin

Comme le nombre de distances cumulées visitées par l'algorithme est le même que DTW classique sans bande de Sakoe-Chiba, nous obtenons un coût de $\mathcal{O}(m \times n)$.

Adaptation de K-Means À chaque itération, l'algorithme K-Means recalcule les barycentres de chaque cluster. Cependant, comme l'algorithme ShiftDTW décale une des séries à chaque itération, ce décalage doit être conservé lors de la mise à jour des barycentres. Chaque série temporelle se voit attribuer un décalage qui lui est propre, selon le barycentre le plus proche et le décalage qui leur est associés.

5 Expérimentations

Dans cette partie, nous allons présenter les résultats obtenus sur 2 jeux de données du domaine, puis sur un échantillon des données métiers. Dans les 2 cas, nous avons utilisé la méthode K-Means++ (Arthur et Vassilvitskii, 2007) pour initialiser les barycentres.

5.1 Jeux de données artificiels

Nous avons d'abord testé notre méthode sur un jeu de données du domaine scientifique afin de s'assurer que notre modèle est pertinent pour répondre à notre problématique de clustering sur données cycliques. Nous avons choisi deux jeux de données, parmi la banque de données UCR time series archive (Dau et al., 2018), qui sont intrinsèquement cyclique car elles sont finies et n'ayant ni début ni fin : BeetleFly et BirdChicken — utilisés pour la première fois par Hills et al. (2014). Ces données représentent les formes de scarabées et de mouches pour le premier, et d'oiseaux et de poulets pour le second, et sont de longueur 512. Nous avons

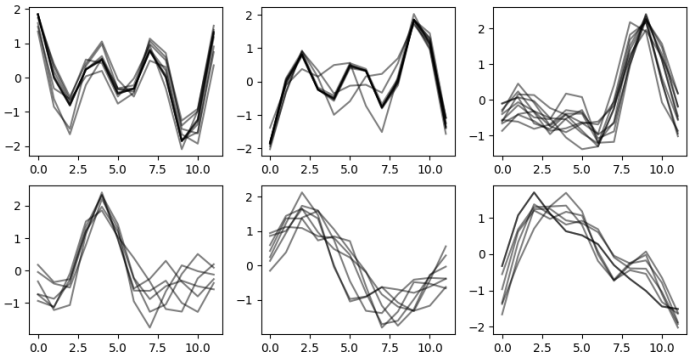


FIG. 4 – Présentation des 6 clusters sélectionnés parmi les 100 clusters calculés avec K-Means euclidien.

testé la distance euclidienne, DTW et Shift DTW avec $r = 4$ avec l'algorithme K-Means, en initialisant 10 fois les barycentres pour ne garder que le meilleur score pour chaque mesure de distance. Ces deux jeux de données contiennent chacun 2 groupes de 10 séries temporelles. Nous présentons dans le tableau 1 les résultats de précision sur les jeux d'apprentissage (train).

	k-means euclidien	k-means DTW	k-means Shift DTW
BeetleFly - train	0.8	0.75	**0.9**
BirdChicken - train	**0.8**	0.7	0.75

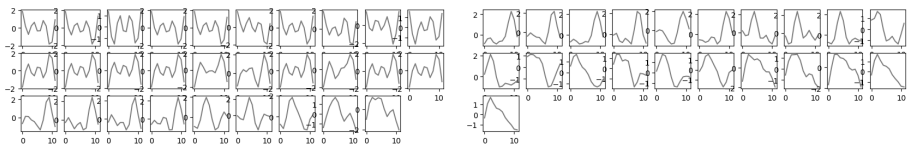
TAB. 1 – Présentation des résultats de classification sur les jeux d'apprentissage BeetleFly et BirdChicken.

On remarque que la mesure Shift DTW est meilleure sur BeetleFly, et est en seconde position sur BirdChicken. Notre mesure montre alors qu'elle est plus pertinente dans certains cas, et qu'elle est meilleure — et donc tout aussi pertinente — que la mesure DTW classique.

5.2 Jeux de données comptables

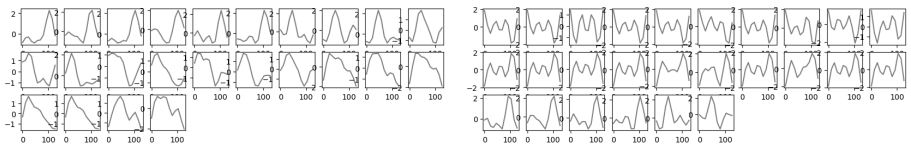
Nous avons ensuite expérimenté notre méthode sur les données métiers. L'objectif ici est de montrer qu'il est possible de regrouper des séries temporelles qui se ressemblent mais qui possèdent un décalage dans le temps. Le décalage doit être assez grand pour que la mesure euclidienne calcule des distances trop grande pour les réunir dans un même cluster K-Mean. Parmi les 1042 séries temporelles, donc chacune d'entre elles représente une et une seule entreprise : nous avons choisi un échantillon de 50 séries temporelles pour ne pas à avoir un étiquetage manuel des séries trop lourd. Pour les choisir, nous avons d'abord appliqué un K-Means euclidien avec 100 clusters, de façon à séparer au mieux ces séries temporelles, et nous avons choisi 6 clusters présenté dans l'image 4 qui nous semblaient pertinents pour notre démonstration. Nous souhaitons, par exemple, que les deux premiers clusters soient réunis pour n'en former qu'un, en tenant compte de la cyclicité avec notre méthode.

En effet, il fallait montrer que certaines séries similaires en tenant compte d'un certain décalage, sont bien regroupées avec notre méthode. Nous avons testé en encodant les séries sur différentes longueurs. Sur des séries temporelles de longueur 12 (une valeur par mois) et un paramètre de décalage à 3, les résultats sont intéressants. Par contre, en augmentant la taille des séries temporelles, et en conservant un petit décalage, les regroupements tendent davantage vers notre objectif : les séries décalées dans le temps se rassemblent mieux. C'est également pour cela que notre méthode a donné de bons résultats sur les jeux précédents. Nous pouvons observer les 2 clusters trouvés sur ces images avec des longueurs 12. Nous pouvons observer les résultats dans les images 5, 6 et 7.



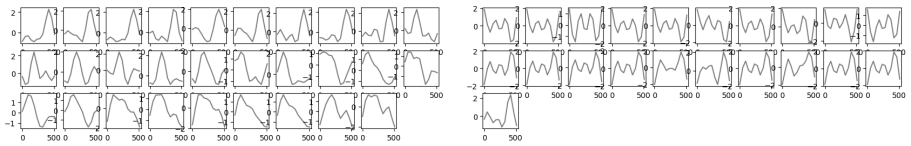
(a) Cluster 1 (b) Cluster 2

FIG. 5 – Résultats du clustering avec les séries temporelles de longueur 12.



(a) Cluster 1 (b) Cluster 2

FIG. 6 – Résultats du clustering avec les séries temporelles de longueur 128.



(a) Cluster 1 (b) Cluster 2

FIG. 7 – Résultats du clustering avec les séries temporelles de longueur 512.

Nous constatons que plus la taille des séries augmente, et plus le premier groupe devient petit et tend à être plus homogène dans la forme des séries qui y sont contenues. On peut toutefois supposer que le deuxième groupe pourrait peut-être se diviser en deux sous-groupes,

mais nous n'avons pas tester avec 3 barycentres. Ces résultats n'auraient pas pu être obtenus avec une distance euclidienne ou bien même avec DTW, et la méthode Cyclic DTW Palazón-González et Marzal (2012) aurait été plus coûteuse en temps de calcul pour parvenir à ces résultats.

6 Conclusion & Perspectives

Nous obtenons de bons résultats en terme de clustering sur ces jeux de données. Nous souhaiterions tester notre méthode sur de plus nombreux jeux de données. Nous aimerions également comparer ces résultats avec ceux de la méthode de Palazón-González et Marzal (2012).

Nous pourrions également tester d'autres algorithmes de clustering tel que Nearest Neighbors Classifier (utilisé également dans Dau et al. (2018)) ou DBSCAN (Ester et al., 1996). Une piste très intéressante est le calcul de barycentre adapté aux mesures DTW proposé par Petitjean et al. (2011). Les auteurs montrent des gains d'inertie intra clusters et globaux très intéressants.

En ce qui concerne son usage sur les données métiers, nous souhaitons améliorer le prétraitement en évaluant, pour chaque entreprise, le meilleur paramétrage à utiliser avec PROPHET pour limiter l'erreur entre les chiffres d'affaire de chaque année et la saisonnalité générée. Il est aussi envisagé d'utiliser notre méthode sur d'autres indicateurs que le chiffres d'affaire, voire de combiner les indicateurs pour travailler en multi-dimensionnel. Cela impose bien sûr d'adapter les méthodes déjà implémentées.

Références

Aghabozorgi, S., A. Seyed Shirkhorshidi, et T. Ying Wah (2015). Time-series clustering – a decade review. *Information Systems 53*, 16–38.

Arthur, D. et S. Vassilvitskii (2007). K-means++ : The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, USA, pp. 1027–1035. Society for Industrial and Applied Mathematics.

Dau, H. A., E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, et Hexagon-ML (2018). The ucr time series classification archive. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

Ester, M., H.-P. Kriegel, J. Sander, et X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press.

Hills, J., J. Lines, E. Baranauskas, J. Mapp, et A. Bagnall (2014). Classification of time series by shapelet transformation. *Data mining and knowledge discovery 28*, 851–881.

Keogh, E. J. et M. J. Pazzani (2000). Scaling up dynamic time warping for datamining applications. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, New York, NY, USA, pp. 285–289. Association for Computing Machinery.

Keogh, E. J. et M. J. Pazzani (2001). Derivative dynamic time warping. In V. Kumar et R. L. Grossman (Eds.), *Proceedings of the First SIAM International Conference on Data Mining, SDM 2001, Chicago, IL, USA, April 5-7, 2001*, pp. 1–11. SIAM.

Maes, M. (1990). On a cyclic string-to-string correction problem. *Information Processing Letters 35*(2), 73–78.

Palazón-González, V. et A. Marzal (2012). On the dynamic time warping of cyclic sequences for shape retrieval. *Image and Vision Computing 30*(12), 978–990.

Petitjean, F., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition 44*(3), 678–693.

Sakoe, H. et S. Chiba (1971). A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics, Budapest*, Volume 3, Budapest, pp. 65–69. Akadémiai Kiadó.

Sakoe, H. et S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*(1), 43–49.

Salvador, S. et P. Chan (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal. 11*(5), 561–580.

Taylor, S. J. et B. Letham (2018). Forecasting at scale. *The American Statistician 72*(1), 37–45.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition 38*(11), 1857–1874.

Summary

The elasticity of the DTW metric provides a more flexible comparison between time series and is used in numerous machine learning domains such as classification or clustering. However, it does not align the measurements at the beginning and end of time series if they have a shift occurring right at the start of one series, with the omitted part appearing at the end of that series. Due to the cyclicity of such series - which lack a definite beginning or end - we rely on the Cyclic DTW approach to propose a less computationally expensive approximation of this calculation method. This approximation will then be employed in conjunction with the K-Means clustering method.

Création d'un référentiel géo-historique d'adresses à partir de sources multiples

Charly Bernard*, Nathalie Abadie*
Julien Perret*,** Bertrand Duménieu**

*LASTIG, Université Gustave Eiffel, IGN/ENSG
{charly.bernard ; nathalie-f.abadie ; julien.perret}@ign.fr,
https://www.umr-lastig.fr/
**CRH, EHESS
bertrand.dumenieu@ehess.fr

Résumé. Énoncés structurés désignant un lieu en s'appuyant sur son contexte géographique, les adresses servent aux individus à localiser un destination précise, un immeuble dans une ville par exemple, avec le moins d'ambiguïté possible. Ces références spatiales indirectes encodent des relations spatiales et hiérarchiques entre un nombre d'entités géographiques parfois important. Prises en masse, elles forment une description d'un territoire à différentes échelles. Lorsqu'il s'agit d'adresses anciennes, elles peuvent aider à reconstituer très finement les structures disparues d'un espace et ont un potentiel particulièrement puissant pour l'étude des dynamiques urbaines sur le temps long. La masse et l'hétérogénéité de ces objets rendent leur exploitation informatisée difficile malgré leur profusion dans les archives numérisées et dans les sources de données structurées sur le Web. Dans cet article, nous présentons une ontologie pour modéliser les adresses dans leur diversité et pour représenter l'évolution temporelle des entités géographiques qui les composent. Nous expliquons aussi comment et avec quelles données nous peuplons l'ontologie créée.

1 Introduction

La disponibilité croissante des sources d'archives numérisées et les progrès récents des approches d'extraction d'informations dans des textes ou des images, permis par l'essor des réseaux de neurones profonds, offrent aujourd'hui la possibilité d'exploiter ces sources pour produire des données structurées sur les divers aspects du passé qu'elles décrivent. Parmi les informations que l'on peut trouver dans les sources d'archives, les adresses constituent un corpus assez important à la fois par son volume et son intérêt applicatif. Premièrement, elles ont une place centrale dans divers types de documents iconographiques comme les plans de villes ou textuels comme les annuaires des habitants et des commerces, les actes d'état civil, les documents notariés, etc. Elles sont donc disponibles en grande quantité avec un taux de recouvrement potentiel entre sources relativement important. Deuxièmement, elles reflètent l'évolution des villes dont elles ont accompagné les changements morphologiques et administratifs

au cours du temps. Par exemple, à Paris, différents systèmes de numérotation des immeubles se sont succédé depuis l'Ancien Régime. Enfin, l'intérêt majeur des adresses réside dans leur double rôle de références spatiales directes, lorsqu'elles sont extraites de cartes et décrites par des coordonnées, et indirectes, quand elles sont représentées sous forme textuelle. En effet, disposer de cette double représentation permet de géocoder les sources d'archives textuelles et de localiser les informations qu'elles renferment à la surface de la Terre.

Afin de structurer et de valoriser les informations relatives aux adresses disponibles sous formes hétérogènes et fragmentaires dans les sources d'archives, nous proposons de construire une base de connaissances géohistorique sur les adresses de Paris. Le choix a été fait de se restreindre à la capitale française, car de nombreuses sources y sont disponibles, sur une période temporelle conséquente et représentent les adresses sous diverses formes. Cette base devra permettre de structurer les connaissances disponibles sur les adresses, de représenter leur évolution temporelle et les sources qui les mentionnent. Cet article présente l'ontologie développée pour représenter les deux premiers aspects des adresses historiques extraites de sources d'archives. Nous l'avons développée en suivant la méthodologie appelée *Simplified Agile Methodology for Ontology Development* dite SAMOD [Per17], qui consiste à séparer un problème de modélisation complexe en sous-problèmes, appelés modelets, plus simples à traiter, valider et tester. Un modelet commence avec un argumentaire en langage naturel qui décrit le sous-problème à traiter, avec un glossaire explicitant les notions et termes importants en jeu. Chaque modelet s'accompagne d'un ensemble de questions informelles de compétence, également en langage naturel, qui représentent les questions auxquelles la base de connaissances doit permettre de répondre. À chaque question est enfin associé un ensemble d'exemple de réponses attendues, qui servent à valider le modelet une fois implémenté. La modélisation de phénomènes géohistoriques est complexe et mobilise des thèmes différents de l'ingénierie des connaissance, la méthode SAMOD permet de les identifier et de les séparer pour faciliter la conception. En outre, cela permet d'opérer en cycles rapides avec une mise à l'épreuve régulière de l'ontologie en cours de construction.

Cet article est organisé de la façon suivante : nous présentons tout d'abord un état de l'art sur la représentation de connaissances sur les adresses et sur l'évolution temporelle des entités géographiques. Puis, nous présentons les deux modelets développés pour représenter des adresses de tous types et leurs évolutions au cours du temps. Enfin, nous peuplons ces modelets à l'aide de données réelles issues de diverses sources pour nous assurer de leur généricité et de leur capacité effective à répondre aux questions de compétences initialement définies.

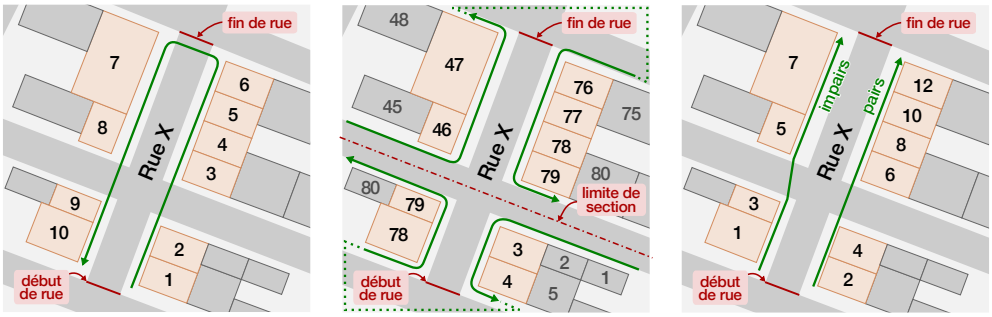
2 État de l'art

2.1 Qu'est-ce qu'une adresse ?

Une adresse est avant tout une référence spatiale indirecte, un énoncé structuré qui désigne un lieu sans ambiguïté [CCD11]. Si la forme et la nature de ce qu'elle désigne varient selon les époques et les lieux, [Mor80] identifie deux catégories. L'adresse administrative, qui définit seulement la position d'un lieu, se distingue de l'adresse cheminement qui décrit un parcours dans l'espace. « *59 rue de Rivoli, 75001 Paris* » et « *maison de la Belle-Jardinière, quai aux Fleurs, coin de la rue de la Cité* » sont respectivement des exemples de ces deux catégories d'adresses. Une adresse est en tout cas une entité composite, faite de références spatiales agen-

cées en hiérarchies et dont les relations spatiales sont éventuellement explicitées [CCD11]. Ces composantes sont des points de repères spatiaux, possiblement informels, connus et partagés par les "adressés" [Dum21]. Dans cet article, une adresse est donc un énoncé structuré d'un cheminement à l'intérieur d'une hiérarchie spatiale, non ambigu au sein de cette hiérarchie, composé d'une suite ordonnée de repères spatiaux dont la conceptualisation et la désignation sont connues et partagées.

2.2 Systèmes d'adressages à Paris depuis le XVIII^{e} siècle



(a) Système de Kreenfelt (b) Système révolutionnaire (c) Système impérial

FIG. 1 – Les trois systèmes de numérotation des habitations utilisés à Paris. Les flèches vertes représentent le sens de numérotation.

Comme dans d'autres capitales européennes, l'adressage des maisons de Paris par des numéros apparaît à la fin du XVIII^{e} siècle. Jusque-là, les bâtiments étaient désignés par un nom, souvent une enseigne, parfois avec des précisions pour le désambiguïser [Pro66], le tout complété de positionnements relatifs, comme la « *Maison du Pied de Griffon […] au coin oriental de la rue du Chantre, […] front sur la rue de Beauvais* » [1]. Trois systèmes de numérotation se succèdent ensuite. Un premier, d'initiative privée, est tenté par l'éditeur de l'*Almanach de Paris* [2], Marin Kreenfelt, en 1779 [Dep16]. Les portes des bâtiments d'une même rue sont numérotées en suivant un parcours en « fer à cheval » (fig. 1a). Sans doute peu usitées dans les faits, ces adresses sont données dans l'Almanach de Paris. L'adressage dit *sectionnaire* apparaît durant la Révolution et ne lui survivra pas longtemps. Nommé d'après la subdivision de Paris en sections, il s'appuie sur ce découpage. Créée dans un but surtout fiscal, cette numérotation n'a jamais été entièrement reconstituée, car elle s'appuyait sur une règle générale inégalement appliquée et sur de nombreuses adaptations locales non documentées [Fle95, Waq18]. Le numérotage débutait à une extrémité d'une section puis suivait un parcours le long des îlots urbains la composant, avec une logique propre à chacune (fig. 1b). Le troisième mode de numérotage, encore en vigueur aujourd'hui, est créé sous le Premier Empire. Il consiste à

1. Adolphe Berty, Henri Legrand, Lazare-Maurice Tisserand, Théodore Vacquer, Camille Platon. Topographie historique du vieux Paris. Région du Louvre et des Tuileries, volume 1. Paris, imprimerie impériale édition, 1866.

2. Marin Kreenfelt. Almanach de Paris, 1re partie contenant la demeure, les noms et qualités des personnes de condition, etc. 2^{e} partie contenant les noms et demeures des principaux artistes, marchands, fabricants, etc. Pour l'année 1789. 1789.

numéroter les unités foncières (parcelle, maisons) de part et d'autre de chaque voie en ordre croissant, numéros impairs à gauche, numéros pairs à droite. Le point de départ de la rue et donc de sa numérotation est, en principe, l'extrémité la plus proche de la Seine (fig. 1c).

2.3 Modéliser les adresses

Si la littérature s'accorde sur la structure générale d'une adresse comme étant un ensemble organisé de points de repère reliés entre eux par des relations spatiales, les adresses postales concentrent l'attention du fait de leur importance administrative. Les propositions de représentation formelles existantes s'occupent essentiellement de ce type d'adresse, notamment dans un but d'harmonisation et de standardisation au sein d'un pays ou d'une région [CCD11, CC07, Zan08, BDS00].

Dans le domaine de la géomatique, la norme ISO 19133 [Iso05] porte uniquement sur des adresses postales. Les modèles à visée générique restent peu adaptés à des modes d'adressages plus hétérogènes. Par exemple, le modèle relationnel proposé par [DF07] décompose les adresses en repères géographiques reliés par des relations spatiales, mais avec une structure hiérarchique systématique : bâtiment, municipalité, région. C'est également le cas de l'ontologie *locn*[3], dont la classe `locn:Address` est une accumulation de `Literal` et non de ressources. Les types de repères y sont limités (rue, unité administrative…) et les relations spatiales ne sont pas indiquées.

Ces propositions, prévues pour le cas des adresses postales contemporaines, ne sont pas immédiatement adaptables aux adresses anciennes de Paris [CDA+18]. Par exemple, le libellé d'adresse « *situé sur le blv. de Clichy, dans la partie comprise entre la place Blanche et la rue Fontaine* » ne peut pas être représenté par ces modèles : aucune municipalité n'est indiquée, il y a strictement plus d'un repère de type Voie et il y a une relation spatiale définie en langage naturel par le terme *entre*. Enfin, aucun de ces modèles n'intègre le fait que les adresses évoluent au cours du temps.

2.4 Représenter l'évolution temporelle des adresses

La représentation de dynamiques spatiales a fait l'objet de plusieurs recherches [Hal12, CTP98, FYZW10, DM11]. Ces travaux ont en commun une question prégnante : qu'est-ce qui définit l'identité d'une entité spatio-temporelle ? Par exemple, à partir de quand considère-t-on qu'une entité est créée ou a disparu ? [Hal12] propose de dissocier l'identité d'une entité de sa réalité physique. Ainsi, une rue n'est pas créée dès lors qu'elle existe sur le terrain, mais dès sa planification. L'évolution temporelle qu'il propose s'appuie sur le modèle de l'*Identity Based Change* [HE00]. Le principe repose sur l'alternance de phases (ou états spatio-temporels) d'une entité durant sa vie. Les changements sont définis selon la transition entre deux phases. [HE00] mentionne trois états que peut avoir une entité géographique : l'existence, la non-existence simple et la non-existence avec une histoire ; [Hal12] en propose cinq. Avec ces approches, on peut supposer que les entités ont des durées de vie infinies qui possèdent divers états d'existence encadrés par des états de non-existence.

Il existe différentes approches pour former un graphe géohistorique. Cependant, les principales sont limitées à des données homogènes à temporalités fixées [Arm88, Dum15, Cos16,

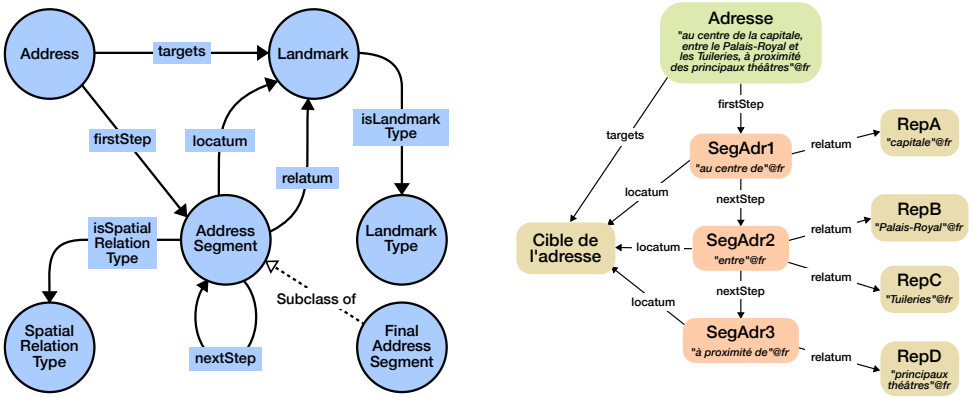
3. `http://www.w3.org/ns/locn`

BVOGD18]. Elles reposent sur la détection des versions d'une entité géographique au sein des sources indépendantes qui décrivent chacune un état spatio-temporel. S'il existe des différences entre les versions successives, des changements en sont déduits. C'est le cas de l'ontologie TSN-Change [BVOGD18] qui permet de tracer l'évolution d'une entité géographique (apparition, disparition, changement de nom, etc.) au sein des différentes versions de la NUTS [4]. Cette nomenclature, publiée chaque année, est un découpage multiniveau du territoire de l'Espace économique européen. L'originalité de l'approche est que les changements peuvent inclure plusieurs entités (fusion, scission…) en les divisant en changements élémentaires qu'on applique au niveau des entités. [BCH+22] propose l'ontologie HHT (Historical Hierarchical Territory) qui applique une approche similaire aux territoires de l'Ancien Régime.

Pour représenter formellement la dimension temporelle de données liées, le W3C propose l'ontologie OWL-time qui permet d'exprimer des faits liés aux ressources avec une approche temporelle. Implémentant l'algèbre temporelle de Allen, elle permet d'associer un événement à un instant ou un intervalle de temps et d'ordonner des faits temporellement [PH06].

3 Proposition

3.1 Adresses



(a) Représentation simplifiée de l'ontologie.

(b) Exemple de modélisation d'une adresse.

FIG. 2 – Modélisation des adresses.

Dans ce modelet [5], une adresse est composée de références à des repères spatiaux, liés par des relations spatiales et qui ensemble pointent vers un lieu cible auquel on peut associer des coordonnées géographiques. Ainsi, nous avons retenu les questions de compétences suivantes : (1) quelles sont les adresses répertoriées le long d'une rue donnée ? (2) quelles sont les coor-

4. Nomenclature des unités territoriales statistiques, voir `https://ec.europa.eu/eurostat/fr/web/nuts/background`

5. `https://github.com/charlybernard/phd-ontologie`

données correspondant à la cible d'une adresse donnée ? (3) quelles sont les adresses localisées dans une zone donnée ? L'ontologie, présentée en figure 2a, comporte trois concepts :

— l'adresse (`Address`), qui correspond au libellé décrivant une localisation ;
— le segment d'adresse (`AddressSegment`) décrivant une partie du cheminement d'une adresse en reliant des repères via une relation spatiale (`SpatialRelationType`) ;
— le repère (`Landmark`) qui décrit une entité géographique. `LandmarkType` définit sa nature (par exemple : unité administrative, voie, numéro d'immeuble, bâtiment…).

Le lieu désigné par l'adresse est dénommé ici `Landmark` et la suite des repères géographiques qui composent l'adresse sont des ressources de type `AddressSegment`. Une ressource de type `Address` est liée via la propriété `firstStep` à une première ressource de type `AddressSegment` qui est elle-même liée via la propriété `nextStep` à une ressource du même type et ainsi de suite. Pour décrire la fin de cette suite, la dernière ressource est de type `FinalAddressSegment` (segment d'adresse final), sous-classe de `AddressSegment`. Une ressource de type `AddressSegment` décrit une étape du cheminement qui lie un repère, qu'on appelle *locatum*, à un ou plusieurs repères qu'on nomme *relatum* [TK11]. Généralement, le locatum est la cible de l'adresse, car chaque étape du cheminement a pour objectif de donner des précisions sur la localisation de la cible. Prenons l'exemple présenté par la figure 2b. L'adresse cible un repère inconnu et a un libellé qu'on peut découper en trois parties : « *au centre de la capitale* », « *entre le Palais-Royal et les Tuileries* » et « *à proximité des principaux théâtres* ». Il y a trois segments et quatre références à des repères géographiques.

3.2 Évolution temporelle

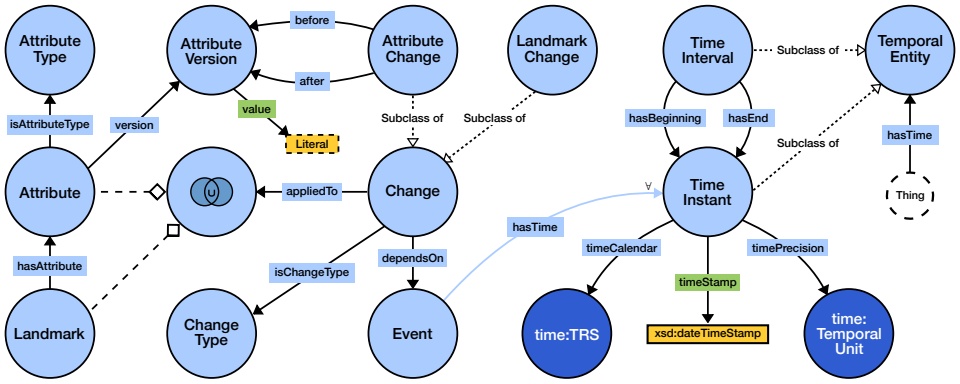


FIG. 3 – Représentation de l'ontologie de l'évolution temporelle des repères géographiques [6].

Pour ce modelet, nous avons identifié les questions de compétences suivantes : (1) quelles sont les voies existant à un instant donné ? (2) sur quel(s) intervalle(s) temporel(s) est valable une adresse de dénomination donnée ? (3) quel est l'historique d'un repère ? Autrement dit,

6. L'élément lié à la classe `Change` par le prédicat `appliedTo` décrit l'union de `Attribute` et de `Landmark`.

quels sont les événements qui lui sont associés ? (4) quels sont les états et les événements qui manquent dans l'historique d'une adresse ? Deux besoins principaux se dégagent de ces questions de compétences : pouvoir retrouver à chaque instant l'état des repères géographiques et représenter les événements ayant eu un effet sur les repères géographiques et leurs propriétés. Pour ce deuxième aspect, on s'inspire des travaux de [BVOGD18] en divisant les événements en changements élémentaires qui explicitent la succession de deux versions. En plus de s'appliquer aux repères, nous avons fait le choix de lier les changements aux attributs des repères. Ainsi, il est possible de déterminer l'évolution d'un attribut en particulier. Pour cela, une propriété liée à un repère est définie via une ressource de type `Attribute` à qui on associe des versions. Dès qu'un changement est appliqué au repère ou à un de ses attributs, on l'indique via une ressource de type `Change` qui décrit à quelle ressource s'applique le changement. [BVOGD18] étudie l'évolution d'entités au sein de *snapshots,* donc les changements décrivent l'évolution entre deux versions successives d'une entité même si aucune différence n'existe. L'approche est discrète, or nous voulons que chaque entité ait un temps de validité qui lui soit propre pour décrire l'état du territoire à chaque instant, ce qui est possible grâce à notre modélisation. La figure 3 présente les quatre concepts principaux de l'ontologie :

1. le changement (`Change`) décrit une opération élémentaire d'un événement. S'il s'applique à un repère géographique, le changement est un `LandmarkChange` et s'il s'applique à un attribut (comme le nom d'une rue), il est un `AttributeChange`. `ChangeType` définit son type (apparition ou disparition d'un repère par exemple) ;
2. l'attribut (`Attribute`) dont le type est donné par `AttributeType` ;
3. la version d'attribut (`AttributeVersion`).

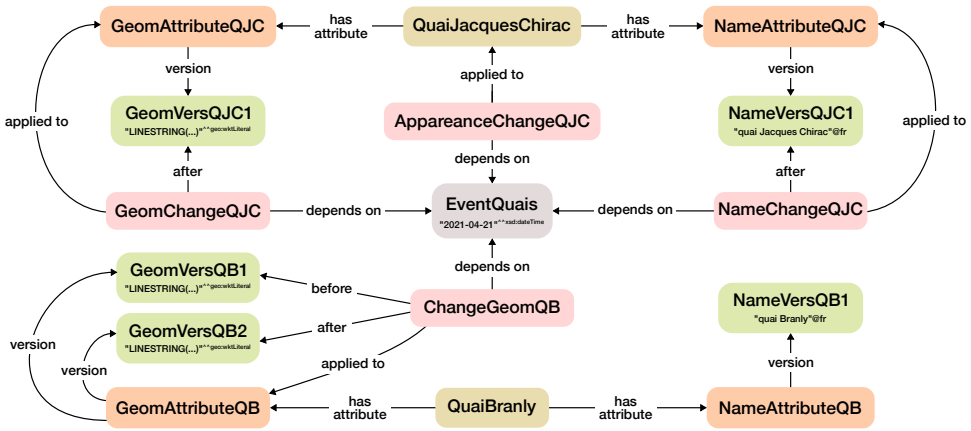


FIG. 4 – Exemple de modélisation de l'évolution conjointe de deux voies : création du *quai Jacques Chirac* par démembrement du *quai Branly*.

Avec cette modélisation, nous pouvons découper des modifications impliquant une ou plusieurs entités géographiques en changements élémentaires. Prenons l'exemple avec l'événe-

ment « création du quai Jacques Chirac sur une portion du quai Branly à la date du 21 avril 2021 ». Nous pouvons le diviser en plusieurs changements comme représentés sur la figure 4 :

— apparition du quai Jacques Chirac (`QuaiJacquesChirac`) ;
— apparition d'une nouvelle version pour `NameAttributeQJC`, attribut de type nom, la version précédente n'existe pas ;
— apparition d'une nouvelle version pour `GeomAttributeQJC`, attribut de type géométrie, la version précédente n'existe pas non plus ;
— apparition d'une nouvelle version pour `GeomAttributeQB`, attribut de type géométrie du quai Branly, la version précédente étant l'emprise spatiale du quai allant de la place de la Résistance au pont d'Iéna.

4 Mise en œuvre et évaluation

La cohérence et le respect des bonnes pratiques de conception des ontologies proposées ont été vérifiés à l'aide du raisonneur HermiT intégré à Protégé [7] [Mus15] ainsi que de l'outil OOPS ! [8] [PVGPSF14].

4.1 Peuplement et validation de l'ontologie des adresses

L'ontologie des adresses est également évaluée en la peuplant avec trois sources de données d'adresses sur Paris ; deux contemporaines : OpenStreetMap (OSM) et la Base Adresse Nationale française (BAN), et une historique : les adresses extraites des annuaires du commerce de Paris du XIX^{e} siècle [ACCD22]. Le processus d'intégration de ces sources comprend trois étapes : (1) identifier les repères et les relations spatiales composant les adresses, (2) structurer chaque adresse selon le modèle proposé, (3) lier les adresses désignant les mêmes cibles.

num	rep	nom_voie	cp	nom_com	lat	lon
5		Boulevard Saint-Martin	75003	Paris 3e	2.361407	48.867933
134		Boulevard Raspail	75006	Paris 6e	2.309998	48.86011
4		Rue Regnard	75006	Paris 6e	2.338073	48.849885
48	bis	Rue de Rivoli	75004	Paris 4e	2.35448	48.85690

TAB. 1 – Extrait d'un fichier de la BAN pour la ville de Paris [9].

La première étape est triviale pour les données OSM et BAN, composées d'adresses postales dont les éléments sont déjà séparés. La structuration des adresses historiques des annuaires du commerce de Paris, plus organiques (figure 2b), est difficile et des outils de *parsing* sur étagère comme LibPostal [10] échouent. Nous avons donc structuré manuellement ces

7. `http://protege.stanford.edu/`
8. `https://oops.linkeddata.es/`
9. Les positions des adresses (fournies par `lat` et `lon`) dans la BAN n'a pas de règle unique définie. Il est indiqué pour chaque adresse la manière dont est fixée cette position (par rapport au point de livraison postal, à la porte d'entrée, au centre de la parcelle, etc).
10. `https://github.com/openvenues/libpostal`

adresses. Les données de chaque source sont ensuite formatées en CSV ou en JSON, éventuellement après une restructuration légère. Par exemple, les données de la BAN (table 1) sont adaptées pour créer un champ `house_number` qui concatène les champs `num` et `rep`.

Un mapping pour chaque source décrit la manière dont les adresses doivent être converties en RDF, conformément à l'ontologie créée : il précise les repères et les relations spatiales entrant en jeu. Nous utilisons le logiciel Ontotext Refine [11] pour construire ces adresses. Le mapping d'un fichier de la BAN permet de créer neuf ressources : une `Address`, quatre `AddressSegment`, quatre `Landmark` (`HouseNumber`, `Thoroughfare`, `PostalCode` et `City`). Nous obtenons ainsi un graphe d'adresses, mais, les adresses étant créées de manière indépendante, les repères entrant en jeu y sont dupliqués : pour les adresses situées le long d'une même rue, on crée autant de ressources différentes pour cette rue que d'adresses. La dernière étape vise donc à éliminer les doublons à l'aide de requêtes SPARQL : les repères de même type, ayant des noms proches et des relations spatiales similaires, sont fusionnés.

Une fois peuplée, nous avons vérifié que l'ontologie permet de répondre aux questions de compétence prédéfinies, transcrites en requêtes SPARQL. Ainsi, la requête pour sélectionner l'ensemble des adresses ciblant un lieu situé le long d'une voie donnée (ici : la rue du Dahomey) renvoie 16 adresses et leurs coordonnées, cartographiées en figure 5.

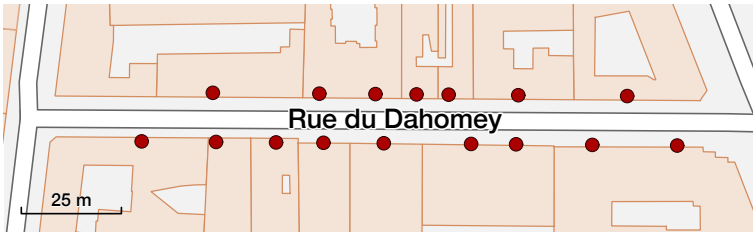


FIG. 5 – Localisation des adresses dont la cible est située le long de la rue du Dahomey.

4.2 Peuplement et validation de l'ontologie des évolutions temporelles

Pour peupler l'ontologie sur l'évolution temporelle des repères géographiques, nous avons utilisé des données de Wikidata sur les voies de Paris. Elles contiennent des informations comme la date de création, une liste des noms officiels avec des dates de début et fin de validité. À partir de ces données, nous créons des repères de type `Thoroughfare`, des attributs de noms avec leurs différentes versions et leurs changements. Wikidata décrit les états des voies. Donc, pour chacun d'entre eux, nous créons deux changements indiquant respectivement le début et la fin de validité de l'état. Cela implique que tous les changements sont indépendants et peuvent être dupliqués. Il faut donc procéder à des fusions via des requêtes SPARQL.

Ce jeu de données se limitant à l'historique des voies, seules certaines questions vues en section 3.2 ont été transcrites en requêtes SPARQL. L'historique d'une entité géographique peut être obtenu soit en requêtant tous les changements opérés sur l'entité, soit en construisant des versions en fonction des changements existants, comme illustré en figure 6. Ici, les états sont déduits des événements et peuvent différer selon les événements qu'on y inclut ou pas.

11. `https://www.ontotext.com/products/ontotext-refine/`

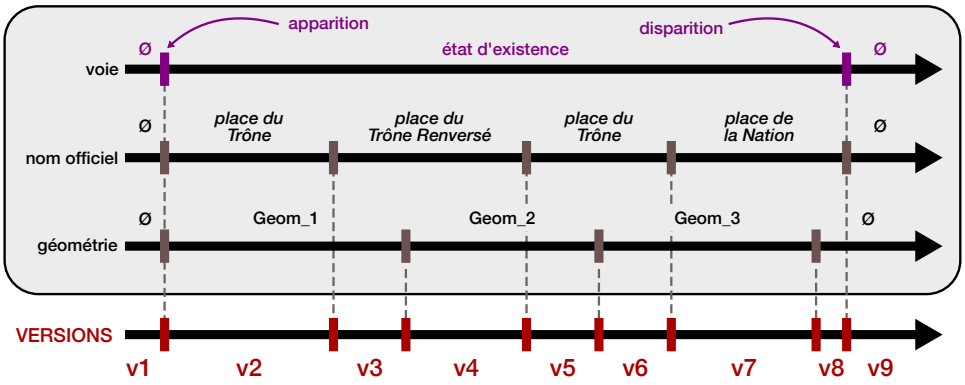


FIG. 6 – Reconstruction des états de la place de la Nation à partir des changements appliqués.

5 Discussion

Dans cet article, nous avons présenté deux ontologies : une pour modéliser les adresses dans leur diversité et une pour représenter l'évolution temporelle des entités géographiques qui les composent. Elle permet la construction cumulative d'un graphe géohistorique à partir de données fragmentaires. Il reste toutefois quelques limites à cette modélisation dont la résolution devra faire l'objet de travaux futurs. Premièrement, nous avons associé aux événements des valeurs temporelles qui sont considérées comme connues et précises. Or, l'imprécision fait partie intégrante des données historiques, et il faut donc prendre en compte cet aspect. Deuxièmement, l'ontologie ne permet pas pour le moment de traiter l'évolution temporelle des relations (notamment spatiales) existant entre les repères géographiques.

En suivant la méthode SAMOD, il reste un modelet à traiter qui concerne les sources et les affirmations qui y sont contenues. Il permettra d'alimenter le graphe à partir de différents fragments d'informations de sources hétérogènes et aussi de gérer les conflits qui peuvent exister entre des sources potentiellement contradictoires.

Enfin, pour valider l'ontologie proposée, nous souhaitons la peupler plus largement, avec différents types de sources de données sur Paris : des sources cartographiques anciennes, dotées de géométries, de type *snapshot*, comme les atlas de Verniquet, de Jacoubet, etc. ; des sources structurées à différentes temporalités comme les *Dénominations caduques des voies de Paris* [12] ou la BAN [13] ; des sources non structurées, également à différentes temporalités, comme des annuaires anciens, ou les descriptions textuelles de l'historique des rues fournies dans Wikipedia. Le graphe de connaissance ainsi peuplé devra permettre d'alimenter un outil de géocodage historique, capable de localiser des intitulés d'adresses dans l'espace parisien et dans leur période d'existence connue.

12. https://opendata.paris.fr/
13. https://www.data.gouv.fr/fr/datasets/base-adresse-nationale/

Références

[ACCD22] N. Abadie, E. Carlinet, J. Chazalon, and B. Duménieu. A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories. In Seiichi Uchida, Elisa Barney, and Véronique Eglin, editors, Document Analysis Systems, Lecture Notes in Computer Science, pages 445–460, Cham, 2022. Springer International Publishing.

[Arm88] Marc P Armstrong. Temporality in spatial databases. GIS/LIS 88 Proceedings : Accessing the world, pages 880–889, 1988.

[BCH+22] Lucas Bourel, William Charles, Nathalie Jane Hernandez, Nathalie Aussenac-Gilles, Victor Gay, and Sébastien Poublanc. Graphes de connaissances pour représenter et analyser l'évolution des territoires en Histoire. In Nicolas Lasolle, Olivier Bruneau, and Jean Lieber, editors, Journées humanités numériques et Web sémantique 2022, pages 23–37, Nancy, France, June 2022. Zenodo. Backup Publisher : Laboratoire AHP-PReST : Archives Henri-Poincaré - Philosophie et Recherches sur les Sciences et les Technologies and LORIA : Laboratoire Lorrain de Recherche en Informatique et ses Applications.

[BDS00] Vinayak Borkar, Kaustubh Deshmukh, and Sunita Sarawagi. Automatically Extracting Structure from Free Text Addresses. IEEE Data Eng. Bull., 23 :27–32, January 2000.

[BVOGD18] Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel, and Hy Dao. Ontologies pour représenter l'évolution des découpages territoriaux statistiques. Revue Internationale de Géomatique, 28(4) :409–437, December 2018.

[CC07] Serena Coetzee and Antony K. Cooper. What is an address in South Africa ? South African Journal of Science, 103(11-12) :449–458, December 2007. Publisher : Academy of Science of South Africa.

[CCD11] Serena Coetzee, Antony K Cooper, and Jeofrey Ditsela. Towards good principles for the design of a national addressing scheme. In 25th International Cartographic Conference (ICC 2011). French Committee of Cartography Paris, France, 2011.

[CDA+18] Rémi Cura, Bertrand Dumenieu, Nathalie Abadie, Benoit Costes, Julien Perret, and Maurizio Gribaudi. Historical Collaborative Geocoding. ISPRS International Journal of Geo-Information, 7(7) :262, July 2018. Number : 7 Publisher : Multidisciplinary Digital Publishing Institute.

[Cos16] Benoît Costes. Vers la construction d'un référentiel géographique ancien : un modèle de graphe agrégé pour intégrer, qualifier et analyser des réseaux géohistoriques. PhD Thesis, Université Paris-Est, November 2016.

[CTP98] Christophe Claramunt, Marius Thériault, and Christine Parent. A qualitative representation of evolving spatial entities in two-dimensional topological spaces. In Innovations In GIS 5, pages 128–142. CRC Press, March 1998.

[Dep16] Thierry Depaulis. Vaugeois, tabletier, connu et inconnu (1re partie). Le Vieux Papier, 41(421) :106–113, July 2016.

[DF07] Clodoveu A. Davis and Frederico T. Fonseca. Assessing the Certainty of Loca-

tions Produced by an Address Geocoding System. GeoInformatica, 11(1) :103–129, March 2007.

[DM11] Géraldine Del Mondo. Un modèle de graphe spatio-temporel pour représenter l'évolution d'entités géographiques. phdthesis, Université de Bretagne occidentale, Brest, October 2011.

[Dum15] Bertrand Dumenieu. Un système d'information géographique pour le suivi d'objets historiques urbains à travers l'espace et le temps. PhD Thesis, École des Hautes Études en Sciences Sociales, December 2015.

[Dum21] Gift Dumedah. Address points of landmarks and paratransit services as a credible reference database for geocoding. Transactions in GIS, 25(2) :1027–1048, 2021.

[Fle95] M. Fleury. Concordance entre la numérotation sectionnaire et la numérotation de type actuel de 2466 maisons de Paris. Commission du Vieux, Rotonde de La Villette, Institut d'histoire de Paris, 1995.

[FYZW10] Y. T. Fan, J. Y. Yang, D. H. Zhu, and K. L. Wei. A time-based integration method of spatio-temporal data at spatial database level. Mathematical and Computer Modelling, 51(11) :1286–1292, June 2010.

[Hal12] Pierre Hallot. L'identité à travers l'espace et le temps. Vers une définition de l'identité et des relations spatio-temporelles entre objets géographiques. PhD Thesis, ULiège - Université de Liège, March 2012.

[HE00] Kathleen Hornsby and Max J. Egenhofer. Identity-based change : a foundation for spatio-temporal knowledge representation. International Journal of Geographical Information Science, 14(3) :207–224, April 2000.

[Iso05] ISO Iso. 19133 : Geographic information-Location-based services. Tracking and navigation. International Organization for Standardization. Technical Report, 2005.

[Mor80] Takashi Morita. Cheminement et adresse. Les Annales de la Recherche Urbaine, 7(1) :27–61, 1980. Publisher : Persée - Portail des revues scientifiques en SHS.

[Mus15] Mark A. Musen. The Protégé Project : A Look Back and a Look Forward. AI matters, 1(4) :4–12, June 2015.

[Per17] Silvio Peroni. A Simplified Agile Methodology for Ontology Development. In Mauro Dragoni, María Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors, OWL : Experiences and Directions – Reasoner Evaluation, Lecture Notes in Computer Science, pages 55–69, Cham, 2017. Springer International Publishing.

[PH06] Feng Pan and Jerry R Hobbs. Time Ontology in OWL. W3C working draft, W3C, 2006.

[Pro66] Jeanne Pronteau. Les Numérotages des maisons de Paris, du XVe siècle à nos jours ... Commission des travaux historiques. Préfecture de la Seine, Service des travaux historiques, 1966.

[PVGPSF14] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. OOPS ! (OntOlogy Pitfall Scanner !) : An On-line Tool for Ontology

Evaluation. International Journal on Semantic Web and Information Systems (IJSWIS), 10(2) :7–34, 2014. Publisher : IGI Global.

[TK11] Thora Tenbrink and Werner Kuhn. A Model of Spatial Reference Frames in Language. In Max Egenhofer, Nicholas Giudice, Reinhard Moratz, and Michael Worboys, editors, Spatial Information Theory, pages 371–390, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[Waq18] Dominique Waquet. Almanachs, cadastre, terriers clés du décodage des numéros sectionnaires d'immeubles à Paris, 1791 - 1805. Annales historiques de la Révolution française, 392(2) :173–183, 2018. Place : Paris Publisher : Armand Colin.

[Zan08] Paul A. Zandbergen. A comparison of address point, parcel and street geocoding techniques. Computers, Environment and Urban Systems, 32(3) :214–232, May 2008.

Summary

Addresses are structured designations of places that are used by individuals to locate a precise destination - a building in a city, for example - with as little ambiguity as possible. These indirect spatial references encode spatial and hierarchical relationships between a large number of geographical entities, and taken in mass form a description of an area at different scales. In the case of old addresses, they can help to reconstruct the vanished structures of an area in great detail, and have a particularly powerful potential for the study of urban dynamics over the long term. The mass and heterogeneity of these objects make them difficult to use in computerised form, despite their abundance in digitised archives and structured data sources on the Web. In this article, we present an ontology for modelling addresses in all their diversity and for representing the temporal evolution of geographical entities. We also explain how and with what data we populate the ontology we have created.

LEODS : un framework pour la publication d'observations satellitaires dans le Web des données

Daniela F. Milon-Flores,* Camille Bernard,* Jérôme Gensel,* Gregory Giuliani**

*Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France,
{daniela.milon-flores,camille.bernard,jerome.gensel}@univ-grenoble-alpes.fr
**University of Geneva Institute for Environmental Sciences, Geneva, Switzerland
gregory.giuliani@unige.ch

Résumé. Les acteurs des territoires (citoyens, politiques, etc.) souhaitent aujourd'hui accéder à des données environnementales ouvertes afin d'analyser l'impact du réchauffement climatique sur leur commune. Les données d'observation de la terre (*Earth Observation* en anglais, (EO data)) provenant de satellites permettent d'analyser la couverture du sol et son évolution au cours des ans. Mais, l'interprétation de ces données est encore réservée à un public de spécialistes, à moins de métadonnées expliquant les indices satellitaires. Également, isolées d'autres données telles que des données socio-économiques et des données sur les politiques environnementales mises en oeuvres, elles ne permettent pas de comprendre les changements environnementaux à l'oeuvre sur les territoires. Nous proposons de recourir aux technologies du Web sémantique (aussi appelé Web des Données Ouvertes et Liées) pour construire un graphe de connaissances intégrant des données EO et d'autres sources de données (politiques publiques, plans d'aménagement des territoires, population, etc.) afin d'observer l'évolution environnementale des territoires et, surtout, d'aider des utilisateurs non experts du domaine à comprendre les tendances observées afin de mieux agir et proposer des politiques environnementales efficaces. Dans cet article, nous présentons la première brique logicielle de ce travail, le *framework LEODS*, pour la publication de données EO dans le Web des données, selon une modélisation spatio-temporelle dans le respect des standards du W3C, garantissant par la suite, l'enrichissement sémantique de ces données.

1 Introduction

La surexploitation des ressources naturelles, telles que les forêts et les mers, ainsi que la pollution de l'air, des sols et de l'eau, conduisent au dérèglement climatique et à la perte de biodiversité. Le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC, *IPCC* en anglais) [1] alerte régulièrement sur les conséquences des actions néfastes de l'Homme sur l'environnement. Afin d'améliorer la prise de décision et mettre en œuvre des politiques environnementales efficaces pour contrer ces tendances négatives, les acteurs des territoires, experts

1. www.ipcc.ch

ou non experts (décideurs politiques, citoyens, chercheurs, journalistes, etc.), ont besoin d'accéder à des données environnementales les aidant à comprendre l'évolution environnementale de leur territoire.

Les programmes de surveillance de la Terre tels que Landsat aux États-Unis [2] et Copernicus en Europe [3] produisent une collection gratuite et ouverte de données satellitaires décrivant la Terre, plus connues sous le nom de *Earth Observation* (*EO data*) dans le domaine. En raison de la quantité massive de ces données EO, la plupart des travaux récents dans le domaine (Lewis et al., 2017; Giuliani et al., 2017; Appel et Pebesma, 2019) proposent d'organiser ces données en Cubes de Données ou EODC (pour *EO data cube*). Un EODC est un cube multidimensionnel généralement composé de quatre dimensions : les bandes spectrales observées, le temps, la latitude et la longitude des points d'observation (Baumann, 2017). Ces EODCs émergent comme une solution technologique pour le stockage, mais aussi la gestion, l'accès et l'analyse de grandes quantités de données EO (Giuliani et al., 2017). Les experts bénéficient ainsi de données prêtes à l'analyse et peuvent produire des indices caractérisant la couverture du sol d'une zone d'intérêt. Par exemple, l'indice de végétation par différence normalisée (*Normalized Difference Vegetation Index*, NDVI), extrait d'une image satellite, donne une mesure de la couverture végétale de la zone observée (Appel et Pebesma, 2019).

Cependant, les EODCs traditionnels, bien que largement utilisés, présentent deux limitations majeures : (1) leur interprétation est encore réservée à un public spécialiste, à moins de métadonnées expliquant les indices satellitaires ; (2) leur relatif isolement d'autres ressources du web (Augustin et al., 2019; Nativi et al., 2017; Giuliani et al., 2019), telles que des données socio-économiques décrivant la population d'un territoire, ses entreprises et industries ou encore, des données traduisant les choix politiques d'aménagement des territoires (plan local d'urbanisme, par exemple), ceci complique grandement l'interprétation des changements observés sur les territoires, et surtout la compréhension des causes de ces changements.

Afin de veiller à l'enrichissement sémantique des données et garantir leur compréhension par le plus grand nombre, mais aussi afin de rompre avec des silos de données, et introduire des liens de causalité entre différents jeux de données, il est pertinent d'adopter les technologies du Web sémantique, aussi appelé Web des Données (Ouvertes) et Liées (LOD Web) pour publier des données conformément aux principes FAIR (*i.e., findable, accessible, interoperable* et *reusable*). En effet, la création de graphes de connaissances (Hogan, 2020) permet de lier dans un même graphe de données différents jeux de données faisant sens analysés ensemble, bien que publiés par des instances différentes. Les données du LOD Web sont publiées sous la forme de triplets RDF (sujet, prédicat et objet). Chaque élément de ces triplets est décri au sein de modèles ontologiques (ou vocabulaires), également publiés dans le Web (Patel et Jain, 2021). Ces ontologies définissent des concepts, facilitant l'interprétation des données pour les humains, mais aussi leur traitement par des machines (Cyganiak et al., 2014) et l'inférence de nouveaux faits et connaissances implicites par raisonnement sur les concepts et leurs relations (Hamdani et al., 2023). Dans ce contexte, l'ontologie RDF Data Cube (QB) est un vocabulaire standard du Web sémantique établi par le W3C sur les bases du vocabulaire SDMX pour la publication de données statistiques. Cette ontologie permet de publier les données selon un modèle multidimensionnel, sous la forme de cubes de données ouvertes et liées (*RDF Data Cubes* en anglais) (Richard Cyganiak et Tennison, 2014). L'approche multi-dimensionnelle

2. `landsat.gsfc.nasa.gov`
3. `www.copernicus.eu/en`

confère au vocabulaire QB un fort potentiel pour la liaison de données hétérogènes mais partageant des dimensions telles que la dimension *temps* et la dimension *espace*. À ce jour, ce vocabulaire est très largement utilisé par les agences statistiques nationales et internationales pour la publication de données socio-économiques, il l'est plus rarement pour la publication de données d'observation de la Terre, bien qu'il faille noter l'initiative conjointe de l'OGC et du W3C pour la publication d'EO raster data via QB (Brizhinev et al., 2017).

Dans le cadre du projet TRACES [4], un projet de recherche collaboratif international entre la France et la Suisse, nous proposons de recourir aux technologies du Web sémantique afin de construire un graphe de connaissances intégrant des données EO associées à d'autres sources de données (politiques publiques, plans d'aménagement des territoires, population, etc.). L'objectif du projet Traces est d'observer l'évolution environnementale des territoires et, surtout, d'aider des utilisateurs experts mais aussi non experts du domaine à comprendre les tendances observées et ainsi, mieux agir et proposer des politiques environnementales efficaces.

Dans cet article, nous présentons la première brique logicielle de cette approche, le *framework* LEODS (pour *Linked Earth Observation Data Series*), pour la publication de données EO dans le LOD Web. LEODS propose une modélisation spatio-temporelle qui adopte les standards RDF Data Cube du W3C, OWL Time [5] du W3C et GeoSPARQL [6] de l'OGC et garantit ainsi l'enrichissement sémantique des données EO. Le framework LEODS prend en entrée des données satellitaires EO et produit en sortie des cubes de données RDF à trois dimensions hiérarchisées : spatiale (communes, départements, pays), temporelle (données journalières, mensuelles, saisonnales, annuelles), et thématique (indices satellitaires, catégories d'indices). Les cubes sont enrichis sémantiquement via des définitions d'indices fournies par les experts du projet TRACES mais aussi via des liens établis vers des graphes de connaissances encyclopédiques tels que Wikidata (Vrandečić et Krötzsch, 2014) ou DBPedia (Auer et al., 2007). La modélisation spatio-temporelle adoptée est un gage de l'enrichissement future des données EO par des données socio-économiques, elles aussi publiées en RDF Data Cube, et partageant des dimensions spatiale et temporelle communes.

La suite de l'article est organisée ainsi : la Section 2 présente les travaux liés à la publication de données EO dans le LOD Web ; la Section 3 introduit le framework LEODS et l'ensemble de la chaîne de traitements qu'il prend en charge. L'étude de cas et les résultats obtenus grâce à LEODS sont présentés et discutés dans la Section 4. Enfin, la Section 5 conclut l'article et présente les perspectives associées à notre travail.

2 Travaux connexes

Les données EO, provenant de satellites, sont des données sensorielles dépourvues de signification sémantique (Augustin et al., 2019). Afin d'interpréter correctement les données EO, il est nécessaire de suivre un processus d'enrichissement sémantique qui améliore les données avec des informations supplémentaires qui fournissent un contexte. Les approches typiques de l'enrichissement sémantique des données EO sont les approches basées sur le contenu, où les caractéristiques extraites des images satellites, par exemple les informations sur les couleurs ou le type d'occupation du sol, sont utilisées pour l'enrichissement, et les approches basées sur

4. http ://traces-anr-fns.imag.fr
5. www.w3.org/TR/owl-time/
6. www.opengis.net/ont/geosparql

l'ontologie, où les modèles sémantiques sont exploités pour représenter et enrichir les données EO dans le LOD Web. Ci-dessous, nous présentons la recherche pertinente à notre proposition.

Dans (Augustin et al., 2019; Sudmanns et al., 2021; Van Der Meer et al.), les auteurs soutiennent que les EODC traditionnels manquent de sémantique et proposent d'ajouter une composante sémantique pour améliorer la recherche d'informations dans les images satellite. Ainsi, en utilisant le logiciel de système expert *Satellite Image Automatic Mapper*, ils enrichissent sémantiquement les EODCs. Cependant, il est important de souligner que la sémantique ajoutée aux EODCs n'est pas publiée en tant que données liées dans le LOD Web, alors que cette information serait très utile aux utilisateurs finaux pour mieux interpréter les données EO. Par la suite, des travaux tels que (Simoes et al., 2021; Datcu et al., 2003) se concentrent également sur l'analyse d'images satellites pour extraire des informations utiles pour les utilisateurs finaux. Par exemple, extraire des caractéristiques de couverture du sol (*e.g.,* montagnes, prairies, eau, etc.) pour améliorer la recherche d'informations dans grand volume de données EO. Cependant, une fois de plus, les résultats de l'analyse sémantique des images ne sont pas publiés en tant que données liées dans le LOD Web.

Pendant notre recherche, nous avons identifié différents vocabulaires/ontologies qui facilitent la publication de données EO sur le LOD Web. Les plus significatifs sont les suivants : (a) *RDF Data Cube Vocabulary* (QB) [7], "un vocabulaire dédié à la publication de données multidimensionnelles. Dans ce modèle de cube lié ouvert, les valeurs appartiennent à un ensemble de données structuré en trois composants : dimensions, mesures et attributs". (b) *RDF Data Cube extensions for spatio-temporal components* (QB4ST) [8], "une extension de RDF Data Cube qui fournit des outils pour définir les caractéristiques spatio-temporelles des dimensions et des mesures." (c) *Sensor, Observation, Sampler, and Actuator* (SOSA) [9] et *Semantic Sensor Network* (SSN) [10], "deux ontologies pour décrire les capteurs [...] ainsi que leurs observations, et les activités d'échantillonnage liées".

Également, en ce qui concerne les approches ontologiques, avec ou sans recours aux ontologies mentionnées précédemment, la première initiative visant à publier des données EO dans le LOD Web a été entreprise par les projets TELEIOS et LEO (Koubarakis et al., 2014). Dans ces deux projets, plusieurs outils, tels que Geotriples (Kyzirakos et al., 2018), et ontologies, telles que *stRDF* (Koubarakis et Kyzirakos, 2010), ont été développés pour publier des données EO dans le LOD Web. Cependant, dans les deux approches, les technologies du LOD Web sont uniquement utilisées pour publier les métadonnées des images satellites sous la forme de données liées. Aucune autre information, telles que, par exemple, les indices satellitaires pouvant être calculés à partir des images, n'est publiée dans le LOD Web. Dans (Brizhinev et al., 2017), une méthode de publication de données EO est proposée. Cette méthode est particulièrement intéressante car elle utilise, entre autres, l'ontologie QB pour représenter les données raster géospatiales à travers ses dimensions, telles que la latitude, la longitude, le temps, etc. Cependant, les données sont publiées par pixel ce qui est très coûteux en termes de stockage de données, alors que, pour les acteurs des territoires, des données agrégées à la commune semblent plus appropriées et significatives. Dans les travaux de (Tran et al., 2020a,b), les auteurs présentent un réseaux d'ontologies du LOD Web pour l'intégration de données calculées à partir des données raster, telles que les indices de couverture du sol. Par la suite, une métho-

7. https://www.w3.org/TR/vocab-data-cube/
8. https://www.w3.org/TR/qb4st/
9. http://www.w3.org/ns/sosa/
10. http://www.w3.org/ns/ssn/

dologie est introduite pour extraire des données pixels des objets des territoires et les relier à un modèle ontologique. Bien que confrontés également à des problèmes de réutilisabilité et d'interopérabilité de leurs données, les auteurs n'ont pas recours à l'ontologie standard QB, alors que ce modèle semble particulièrement adapté à la modélisation spatio-temporelle et propice à faciliter les connexions à d'autres données telles que des données socio-économiques. En effet, QB a déjà été utilisé pour structurer des données de différentes natures dans le LOD Web : des données médicales (Rodriguez et Hogan, 2021; Casey et al., 2022) et historiques (Bayerl et Granitzer, 2015). En particulier, les travaux de Lefort et al. (2012); Ayadi et al. (2022) décrivent une approche pour représenter les données météorologiques, dérivées des EOs, pour surveiller la variabilité du climat et capturer son comportement. Dans les deux approches, le vocabulaire QB est utilisé pour créer des segments spatio-temporels d'observations météorologiques enrichis d'attributs statistiques. De plus, les méthodes exploitent l'ontologie du SSN pour effectuer une description plus personnalisée des observations, telles que les caractéristiques liées aux senseurs. Bien que cette procédure contextualise les EO de manière granulaire, le résultat est un cube de données limité à la représentation des seules données météorologiques. Nous pensons que cette partie de leur approche diverge de notre objectif, car les parties prenantes non expertes ne sont pas concernées par une description aussi basse des observations et l'absence de ces données n'affecte pas l'interprétation de l'évolution de l'environnement.

Dans notre travail, nous nous intéressons à l'ouverture des données EO à un large public en réutilisant des modèles sémantiques standard, incluant le cube de données RDF. Contrairement à la plupart des travaux connexes décrits précédemment, nous ne nous concentrons ni sur la simple représentation des métadonnées comme LOD, ni sur le traitement des données EO au niveau matriciel. Nous nous concentrons plutôt sur les données EO agrégées au niveau administratif le plus bas, afin d'être aussi proches que possible des parties prenantes, expertes ou non, qui s'intéressent aux aspects environnementaux de leurs municipalités. En outre, nous visons un niveau plus élevé d'enrichissement sémantique au-delà du domaine de l'EO afin de contextualiser les données avec les ressources Web disponibles de différentes natures, telles que les données socio-économiques. Par conséquent, un framework spécifique est nécessaire pour décrire comment préparer, modéliser, publier et explorer les données spatio-temporelles EO sur le LOD Web. L'approche devrait adopter une modélisation multidimensionnelle suffisamment générique (a) pour être applicable à toute donnée EO mesurée globalement et (b) pour assurer l'intégration avec des cubes de données publiés dans le LOD Web. Dans la section suivante, nous décrivons notre proposition, le framework LEODS.

3 Le framework LEODS

Dans cette section, nous présentons le *framework LEODS* qui permet la transformation de séries temporelles brutes de données EO en cubes de données RDF, agrégées par commune. Comme illustré par la Figure 1, LEODS couvre une chaîne de traitements qui consiste à : (1) agréger les données EO, initialement au niveau du pixel, vers le niveau communal, (2) concevoir un modèle dimensionnel qui structure les cubes de données, (3) instancier ce modèle et publier les données sous forme de cubes de données RDF dans le LOD Web, et (4) fournir aux utilisateurs des requêtes SPARQL préconstruites pour explorer les données. Les sous-sections suivantes expliquent chaque étape du framework.

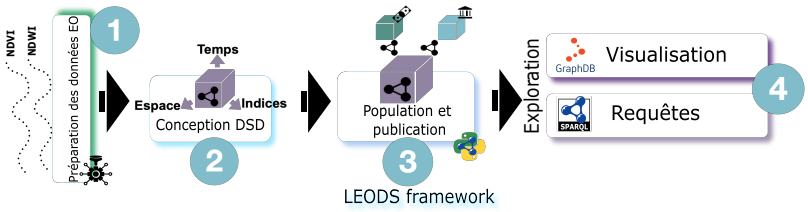


FIG. 1 – *Pipeline du framework LEODS.*

3.1 Étape 1 : Préparation des données

La phase initiale gérée par notre framework vise à agréger les données EO initialement au niveau du pixel pour (1) répondre aux besoins des acteurs territoriaux et (2) manipuler des données structurées multidimensionnelles. Ce processus commence par l'acquisition de données satellitaires pour la zone étudiée et la période d'observation souhaitée, suivie d'étapes de prétraitement essentielles telles que la correction atmosphérique et le débruitage. Les pixels traités sont ensuite agrégés au niveau communal et des indices dérivés avec diverses agrégations temporelles (par exemple, saisonnières, annuelles) peuvent alors être calculés. De plus, des méthodes statistiques zonales, telles que des calculs de moyenne et d'écart-type, sont appliquées afin de résumer les données relatives à chaque indice, sur une période de temps (passage de données journalières à des données mensuelles, saisonnales, ou annuelles). Enfin, les données agrégées obtenues sont extraites sous la forme de séries chronologiques numériques stockées dans des fichiers tabulaires.

3.2 Étape 2 : Modélisation de la structure du cube

La deuxième étape du framework LEODS consiste à modéliser la structure des cubes de données EO-RDF. Cette étape est primordiales et consiste à associer aux valeurs de la série temporelle EO, obtenue lors de la procédure précédente, des composants du cube de données RDF, tels que : (1) Les *observations* sont les valeurs observées dans le jeu de données, jeu de données qui suit une structure bien définie. Cette structure est connue sous le nom de définition de la structure de données (DSD), et chaque jeu de données a exactement une DSD associée. (2) Les observations d'un jeu de données peuvent être organisées selon des *dimensions*. En raison de la nature des données EO, l'espace et le temps sont deux dimensions standard qui doivent être incluses dans le cube de données. En plus des dimensions spatio-temporelles, on peut également trouver des dimensions spécifiquement liées au sujet de le jeu de données. Par exemple, dans le projet TRACES, nous nous intéressons aux aspects environnementaux. Par conséquent, une dimension qui devrait également faire partie du cube sont les Indices de couverture du sol dérivés des images satellites. À son tour, chaque dimension peut être définie de manière hiérarchique pour une meilleure modélisation. (3) Un *attribut* est l'unité qui donne du contexte à une observation. Par exemple, la valeur 50 prend du sens lorsqu'elle est associée à l'attribut hectares (HA). (4) Une *mesure* fait référence au phénomène spécifique sous observation et est étroitement liée à l'attribut. Dans le domaine des données EO, des mesures typiques comprennent la mesure de la couverture sol ou la température de surface.

Indices de couverture du sol				
Index des familles	*Index var*	*Nom*	*Nombre total d'indices*	*Statistiques zonales*
LIS	NDSI	Normalized Difference Snow Index	20	Moyenne (sans unité et $°C$), écart-type (sans unité et $°C$), et qualité des données (%)
	NDVI	Normalized Difference Vegetation Index		
	WRI	Water Ratio Index		
	NDBI	Normalized Difference Built-Up Index		
	...			
LST	st	Surface Temperature	1	
CLC	clc-11	Urban fabric	15	Couverture du sol(HA et %))
	clc-22	Permanent crops		
	clc-31	Forests		
	...			

TAB. 1 – *Trois familles d'indices composent nos études de cas : LIS, LST et CLC.*

3.3 Étape 3 : Population et publication de cubes de données EO-RDF

Après avoir modélisé la structure des cubes de données RDF, l'étape suivante consiste en la transformation semi-automatique des séries temporelles EO brutes en cubes de données RDF. Ce processus peut être facilité en utilisant des outils tels que la bibliothèque RDFlib [11], le langage de mapping RML [12], et Tarql [13]. Nous proposons d'utiliser la bibliothèque RDFlib pour instancier le modèle conceptuel des cubes (la DSD) en triplets RDF à l'aide de Python, du fait de la simplicité de la syntaxe de cette librairie. Une fois la DSD instanciée, nous suggérons d'utiliser Tarql pour mettre en œuvre des scripts qui importent automatiquement les observations à partir de fichiers tabulaires (par exemple, CSV et TSV) vers des fichiers RDF (par exemple, Turtle). En fin de compte, les fichiers DSD et d'observation sont combinés en un seul fichier décrivant des cubes de données EO-RDF. De plus, il est fortement recommandé rendre disponibles les données EO-RDF via des services web spécialisés, tels que les points d'accès SPARQL, qui permettent une interrogation et une extraction faciles des données stockées dans les graphes RDF.

3.4 Étape 4 : Exploration des cubes de données EO-RDF

Enfin, une fois que les cubes de données EO-RDF sont accessibles, il devient crucial d'initier la phase d'exploration, démontrant comment extraire des informations significatives des ces cubes. Nous avons utilisé ici l'outil GraphDB [14], qui, en plus de stocker des triplets, comporte un module de visualisation des graphes de données liées. Cette visualisation est essentielle pour découvrir et naviguer à travers les composants du cube. GraphDB prend en charge les requêtes SPARQL qui sont essentielles pour extraire et manipuler les cubes de données EO-RDF. Nous proposons de mettre à disposition des utilisateurs des requêtes SPARQL préconstruites, facilitant ainsi leur exploration et leur interprétation.

4 Résultats et discussion

Les partenaires du projet TRACES sont les créateurs du Swiss Data Cube (SDC) [15]. Le SDC est un EODC qui fournit des données prêtes à l'analyse sur le territoire de la Suisse et

11. https://rdflib.readthedocs.io/en/stable/
12. https://rml.io/specs/rml/
13. https://tarql.github.io/
14. https ://graphdb.ontotext.com/
15. https://www.swissdatacube.org/

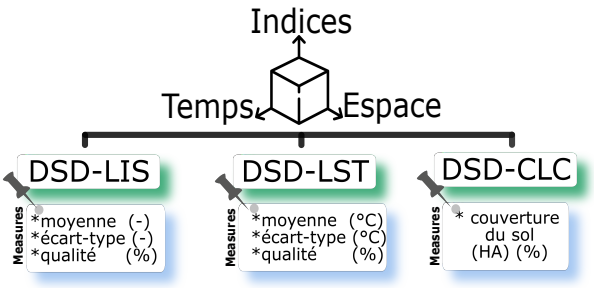


FIG. 2 – *Conception finale de la structure des trois cubes de données EO-RDF.*

une partie de la France, depuis 1984 (Giuliani et al., 2017). Nous commençons par sélectionner trois études de cas du SDC, à savoir les communautés (ensemble de communes) d'Evian en France, de Fribourg en Suisse et de Grand Genève situé à la frontière des deux pays. En résultat de l'agrégation des données au niveau communal, un total de 373 communes et trois familles d'indices ont été obtenues : les indices de caractérisation de la couverture du sol Landsat (LIS), la température de surface du sol Landsat (LST) et la couverture du sol selon la nomenclature Corine (CLC). (Voir Tableau 1) Les indices LIS et LST sont disponibles sous forme de données saisonnières (3 observations par an) pour chaque année sur la période allant de 1985 à 2022, tandis que CLC comporte 5 observations correspondant aux années 1990, 2000, 2006, 2012 et 2018. De plus, des statistiques zonales telles que la moyenne, l'écart-type, la qualité des données et le pourcentage de l'occupation du sol ont également été calculées pour chaque indice.

La deuxième phase du framework conduit à la modélisation de la structure du cube de données EO-RDF. Comme le montre la Figure 2, parmi les trois études de cas, nous identifions trois dimensions : le temps, l'espace et les indices, où "indices" est une dimension thématique liée au sujet de notre jeu de données. En ce qui concerne l'identification des mesures et des attributs, pour CLC, nous identifions facilement la mesure de la couverture du sol avec deux unités (HA et %). Peour LIS et LST, nous avons identifié les mêmes trois mesures, c'est-à-dire la moyenne, l'écart-type et la qualité des données, avec des attributs différents tels que $°C$, $\%$ et attributs sans unité (-). Par conséquent, trois DSD ont été instanciées pour préserver les différentes unités dans le même cube. De plus, nous considérons une structure hiérarchique pour chacune des dimensions. Par exemple, il est établi qu'une commune appartient à un pays (*par exemple,* la France et la Suisse) et que des indices de bas niveau tels que le NDVI appartiennent au domaine d'application de la végétation (voir le Tableau 1). Ensuite, des métadonnées et des données liées ont été ajoutées aux composants du cube. À cet égard, une contribution importante est constituée par les métadonnées des indices. Il n'y a pas beaucoup d'informations sur des indices spécifiques tels que le NDBI dans le Web, donc des descriptions, des formules de calcul et des citations bibliographiques ont été ajoutées. Nous avons effectué un processus similaire pour les communes en les connectant à des ressources officielles de données liées, par exemple, les communes en Suisse ont été connectées à son service de données liées officiel [16]. Enfin, en utilisant les scripts Python et Tarql mis en œuvre, nous avons obtenu trois cubes de

16. https ://geo.ld.admin.ch/

```
SELECT ?code ?geometry
WHERE {
?obs rdf:type qb:Observation ;
    :dimensionArea  ?area ;
    :dimensionTime  ?time ;
    :dimensionIndice traces-codelist:
                        NDVI ;
    :measureMeanUnitless ?mean ;
?area sett:studyArea traces-geo:
                    Fribourg ;
tsn:hasIdentifier ?code ;
owl:sameAs ?bounderies .
?time time:year  "2022"^^xsd:gYear ;
sett:seasonOfYear sett:Spring .

SERVICE <https://geo.ld.admin.ch/query>
{
    ?bounderies geo:hasGeometry ?
                        coordinates .
    ?coordinates geo:asWKT ?geometry .
} }
ORDER BY DESC(?mean)
LIMIT 3
```

FIG. 3 – *Requête spatio-temporel.*

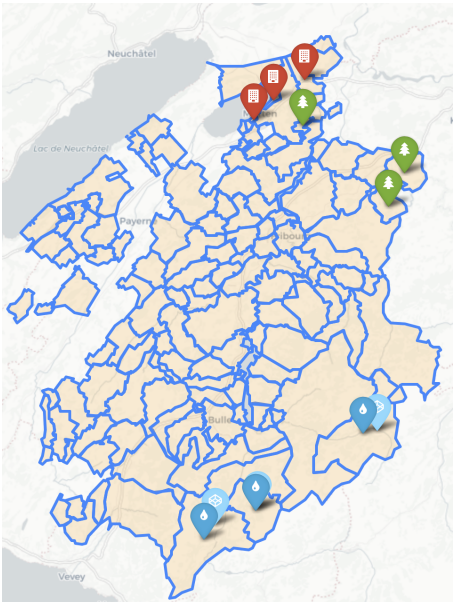


FIG. 4 – *Communes pertinentes pour le canton de Fribourg.*

données EO-RDF contenant environ 21 millions de triples qui représentent les observations concernant les trois études de cas. Actuellement, les utilisateurs peuvent accéder aux cubes via un point d'accès SPARQL.

4.1 Exploration des données

Les requêtes SPARQL servent à extraire des informations des cubes de données RDF générés. Dans le cadre de notre méthodologie, nous avons l'intention de fournir aux parties prenantes des requêtes SPARQL prédéfinies qui facilitent l'exploration des données liées. Bien que notre attention actuelle dans cette étude se concentre sur les requêtes spatio-temporelles, nous avons mis en place différents types de requêtes actuellement accessibles dans un dépôt GitHub [17]. Le code comprend des requêtes de base pour récupérer les composants du cube tels que les dimensions, les mesures et les attributs, ainsi que des requêtes plus complexes exploitant des composants hiérarchiques pour des opérations OLAP [18] telles que le *Drill-down* et le *Roll-up*. Sur la Figure 3, nous présentons une requête spécifique pour illustrer la polyvalence de notre structure cubique finale qui manipule des dimensions standard et thématiques et extrait des informations géographiques non présentes à l'origine dans le jeu de données. Plus précisément, en utilisant nos cubes RDF, la requête obtient les trois principales communes présentant les valeurs moyennes de NDVI les plus élevées dans l'étude de cas de Fribourg en 2022. Ensuite, parce que chaque commune est connectée à sa source de données liées officielle

17. https://github.com/DanielaFe7-personal/QBSparqlQueries
18. *Online Analytical Processing* (OLAP) en anglais.

respective, nous avons extrait les coordonnées spatiales des limites des communes concernées. Le lecteur doit noter que ces informations ne figuraient pas dans le jeu de données d'origine, ce qui est l'une des principales caractéristiques du Web sémantique. Les informations supplémentaires pouvant être obtenues incluent le nombre d'habitants, le pays de provenance, etc. Par la suite, en utilisant la bibliothèque Python Folium [19], il est possible de représenter les coordonnées de ces communes sur une carte. Pour le graphique final dans la Figure 4, nous répétons la requête précédente en faisant varier la valeur de l'indice. Plus précisément, nous affichons les trois communes avec les valeurs moyennes de NDVI les plus élevées en vert (Végétation), NDBI en rouge (Urbanisation), NDWI en bleu (Eau), et NDSI en bleu clair (Neige). Ainsi, en exploitant les requêtes et les visualisations, nous pouvons obtenir une perspective graphique d'une commune donnée. Les premiers résultats suggèrent que les zones plus fréquentées par les résidents se trouvent dans le nord de Fribourg où les indices d'urbanisation et de végétation prédominent, tandis que la zone montagneuse se trouve au sud à cause de la prédominance des indices de neige et d'eau. La requête peut également être répétée pour différentes années, ce qui permet une comparaison temporelle des résultats.

5 Conclusions

Dans cet article, nous avons proposé un framework pour la publication d'observations satellitaires dans le LOD Web. Grâce à LEODS, les séries temporelles EO sont stockées dans des cubes de données EO-RDF qui ne sont plus isolés, mais enrichis par des métadonnées et des liens vers diverses ressources dans le LOD Web. Avec LEODS, nous espérons que les experts et les non-experts pourront bénéficier de notre approche. Par exemple, les trois cubes de données EO-RDF produits intègrent des informations pertinentes sur les aspects environnementaux de communes sélectionnées en Suisse et en France. Les experts, les citoyens, les décideurs politiques et les associations intéressés ont maintenant à leur disposition des données EO RDF au niveau communal. Finalement, dans le cadre de travaux futurs, nous souhaitons connecter nos cubes RDF avec d'autres données du LOD Web, telles que des données socio-économiques basées sur les dimensions spatio-temporelles implémentées dans nos cubes EO. Il est également nécessaire d'implémenter un module pour gérer la scalabilité lors de l'intégration de nouvelles données dans le cube. De plus, le lecteur doit noter que LEODS est une première étape dans la création d'un graphe de connaissances représentant les trajectoires environnementales des municipalités.

Références

Appel, M. et E. Pebesma (2019). On-demand processing of data cubes from satellite image collections with the gdalcubes library. *Data 4*(3), 92.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, et Z. Ives (2007). Dbpedia : A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, et P. Cudré-Mauroux (Eds.), *The Semantic Web*, Berlin, Heidelberg, pp. 722–735. Springer Berlin Heidelberg.

19. `https://python-visualization.github.io/folium/latest/#`

Augustin, H., M. Sudmanns, D. Tiede, S. Lang, et A. Baraldi (2019). Semantic earth observation data cubes. *Data 4*(3), 102.

Ayadi, N. Y., C. Faron, F. Michel, F. Gandon, et O. Corby (2022). A model for meteorological knowledge graphs : Application to météo-france data. In *ICWE 2022-22nd International Conference on Web Engineering*.

Baumann, P. (2017). The datacube manifesto.

Bayerl, S. et M. Granitzer (2015). Data-transformation on historical data using the rdf data cube vocabulary. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, pp. 1–8.

Brizhinev, D., S. Toyer, K. Taylor, et Z. Zhang (2017). Publishing and using earth observation data with the rdf data cube and the discrete global grid system. *W3C Working Group Note and OGC Discussion Paper W3C 20170928*, 16–125.

Casey, S., P. Doody, et A. Shields (2022). An ontology-based system for cancer registry data. In *2022 33rd Irish Signals and Systems Conference (ISSC)*, pp. 1–6.

Cyganiak, R., D. Wood, et M. Lanthaler (2014). RDF 1.1 Concepts and Abstract Syntax.

Datcu, M., H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P. Marchetti, et S. D'Elia (2003). Information mining in remote sensing image archives : system concepts. *41*(12), 2923–2936.

Giuliani, G., B. Chatenoux, A. De Bono, D. Rodila, J.-P. Richard, K. Allenbach, H. Dao, et P. Peduzzi (2017). Building an earth observations data cube : lessons learned from the swiss data cube (sdc) on generating analysis ready data (ard). *Big Earth Data 1*(1-2), 100–117.

Giuliani, G., J. Masó, P. Mazzetti, S. Nativi, et A. Zabala (2019). Paving the way to increased interoperability of earth observations data cubes. *Data 4*(3).

Hamdani, Y., G. Xiao, L. Ding, et D. Calvanese (2023). An ontology-based framework for geospatial integration and querying of raster data cube using virtual knowledge graphs. *ISPRS International Journal of Geo-Information 12*(9), 375.

Hogan, A. (2020). The semantic web : Two decades on. *Semantic Web 11*(1), 169–185.

Koubarakis, M. et K. Kyzirakos (2010). Modeling and querying metadata in the semantic sensor web : The model strdf and the query language stsparql. In *Extended Semantic Web Conference*, pp. 425–439. Springer.

Koubarakis, M., K. Kyzirakos, C. Nikolaou, G. Garbis, K. Bereta, P. Smeros, S. Gianakopoulou, K. Dogani, M. Karpathiotaki, et I. Vlachopoulos (2014). Linked earth observation data : The projects teleios and leo. In *Proceedings of the Linking Geospatial Data Conference*.

Kyzirakos, K., D. Savva, I. Vlachopoulos, A. Vasileiou, N. Karalis, M. Koubarakis, et S. Manegold (2018). Geotriples : Transforming geospatial data into rdf graphs using r2rml and rml mappings. *Journal of Web Semantics 52*, 16–32.

Lefort, L., J. Bobruk, A. Haller, K. Taylor, et A. Woolf (2012). A linked sensor data cube for a 100 year homogenised daily temperature dataset. In *SSN*, pp. 1–16.

Lewis, A., S. Oliver, L. Lymburner, B. Evans, L. Wyborn, N. Mueller, G. Raevksi, J. Hooke, R. Woodcock, J. Sixsmith, et al. (2017). The australian geoscience data cube—foundations and lessons learned. *Remote Sensing of Environment 202*, 276–292.

Nativi, S., P. Mazzetti, et M. Craglia (2017). A view-based model of data-cube to support big earth data systems interoperability. *Big Earth Data 1*(1-2), 75–99.

Patel, A. et S. Jain (2021). Present and future of semantic web technologies : a research statement. *International Journal of Computers and Applications 43*(5), 413–422.

Richard Cyganiak, D. R. et J. Tennison (2014). The rdf data cube vocabulary. W3c recommendation, W3C. https ://www.w3.org/TR/vocab-data-cube/.

Rodriguez, T. N. et A. Hogan (2021). Covidcube : An RDF data cube for exploring among-country COVID-19 correlations. Volume 2980 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Simoes, R., G. Camara, G. Queiroz, F. Souza, P. R. Andrade, L. Santos, A. Carvalho, et K. Ferreira (2021). Satellite image time series analysis for big earth observation data. *13*(13), 2428. Publisher : MDPI.

Sudmanns, M., H. Augustin, L. van der Meer, A. Baraldi, et D. Tiede (2021). The austrian semantic eo data cube infrastructure. *Remote Sensing 13*(23), 4807.

Tran, B.-H., N. Aussenac-Gilles, C. Comparot, et C. Trojahn (2020a). An approach for integrating earth observation, change detection and contextual data for semantic search. pp. 3115–3118. IEEE.

Tran, B.-H., N. Aussenac-Gilles, C. Comparot, et C. Trojahn (2020b). Semantic integration of raster data for earth observation : An rdf dataset of territorial unit versions with their land cover. *ISPRS International Journal of Geo-Information 9*(9), 503.

Van Der Meer, L., M. Sudmanns, H. Augustin, A. Baraldi, et D. Tiede. SEMANTIC QUERYING IN EARTH OBSERVATION DATA CUBES. *XLVIII-4/W1-2022*, 503–510.

Vrandečić, D. et M. Krötzsch (2014). Wikidata : A Free Collaborative Knowledgebase. *57*(10), 78–85.

Summary

Non-expert stakeholders (e.g. citizens, policy-makers, etc.) aspire to access open environmental data to analyze the impact of climate change on their municipalities. Earth observation (EO) data from satellites allow the analysis of land cover and its evolution over the years. However, the interpretation of these data must be performed by specialists, unless they are accompanied by metadata explaining the satellite indices. Furthermore, when isolated from other data, such as socioeconomic data and data related to environmental policy implementation, EO data alone do not provide a comprehensive understanding of the environmental changes occurring in municipalities. We propose to leverage Semantic Web technologies (also calles Linked Open Data (LOD) Web) to build a knowledge graph that integrates EO data with other data sources (public policies, municipalities' development plans, population, etc.) to observe the environmental evolution of territories. Importantly, this approach aims to help non-expert users understand the observed trends to better inform actions and propose effective environmental policies. In this paper, we present the first software component of this work, the LEODS framework, focus on publishing EO data in the LOD Web following a spatio-temporal modeling approach compliant with W3C standards and ensuring the subsequent semantic enrichment of EO data.

Apprentissage interprétable de la criminalité en France (2012-2021)

Nida Meddouri*, David Beserra*

* Laboratoire de Recherche de l'EPITA (LRE),
Le Kremlin-Bicêtre, France.
firstname.lastname@epita.fr,

L'activité criminelle en France a connu une évolution significative au cours des deux dernières décennies, marquée par la recrudescence des actes de malveillance, notamment liés aux mouvements sociaux et syndicaux, aux émeutes, ainsi qu'au terrorisme. Dans ce contexte difficile, l'utilisation de techniques issues de l'intelligence artificielle pourrait offrir de nombreuses perspectives pour renforcer la sûreté publique et privée en France. Un exemple de cette approche est l'analyse spatio-temporelle des données de criminalité, déjà couronnée de succès au Brésil (Da Silva et al., 2020), au Proche-Orient (Tolan et al., 2015), et dans d'autres pays. Dans le cadre de ce travail, nous explorons la possibilité d'appliquer cette approche au contexte français.

1 Données de criminalité en France (2012-2021)

Les données de criminalité en France, disponibles en ligne depuis le 9 octobre 2015 sous une licence ouverte (Open Licence) [1], offrent une vue complète des statistiques de crimes et délits enregistrés par les services nationaux de police et de gendarmerie. Ces données couvrent la période de 2012 à 2021 et englobent la France métropolitaine, les *Départements et Régions d'Outre-Mer (DROM)* ainsi que les *Collectivités d'Outre-Mer (COM)*.

Les informations sont déclinées en fonction des différents services de police et de gendarmerie qui les ont enregistrées, et elles proviennent de l'outil historique de mesure de l'activité des services, connu sous le nom "Etat 4001", en place depuis 1972. Ce dispositif statistique utilise une nomenclature comprenant 107 catégories d'infractions, avec 107 index au total, dont 4 ne sont pas utilisés. Cette base de données constitue ainsi une ressource essentielle pour comprendre et analyser l'évolution de la criminalité en France au fil du temps, et nous considérons que ces données peuvent être utilisées dans des analyses spatio-temporelles afin d'aider à "prévoir" et "interpreter" l'occurrence de certains types de crimes en France.

2 Analyse et pré-traitement des données criminels en France

Avant d'utiliser ces données, il est nécessaire de les prétraiter afin de regrouper les statistiques fournies par les services de la police nationale avec celles de la gendarmerie nationale

1. https ://www.data.gouv.fr/fr/datasets/crimes-et-delits-enregistres-par-les-services-de-gendarmerie-et-de-police-depuis-2012/information

par année. Ensuite, étant donné que ces statistiques proviennent de 372 *Compagnies de Gendarmerie Départementale (CGD)* et 828 *Circonscriptions de Sécurité Publique (CSP)*, elles ont été regroupées par département.

Ce choix s'explique par deux raisons principales : tout d'abord, les périmètres de sécurité en France (*CGD* et *CSP*) ont été mis à jour depuis 2011, impliquant des ajouts, regroupements, divisions, etc. Ensuite, la politique sécuritaire en France est décidée au niveau central avant d'être appliquée dans un premier temps au niveau des 101 départements, puis dans un deuxième temps au niveau des 1200 périmètres de sécurité locaux et régionaux (CGD et CSP).

Étant donné que le découpage du territoire français en 101 départements n'a pas changé depuis 2011, nous avons choisi de regrouper les statistiques par département. Enfin, ces données seront étiquetées en fonction de la survenue ou non d'une émeute, d'un attentat, d'un attentat déjoué ou d'aucun de ces événements.

3 Apprentissage interprétable des comportements criminels

Ainsi, l'objectif de ce travail est d'appréhender l'évolution du comportement criminel à partir des données étiquetées. Nous proposons d'apprendre à partir des données correspondant à chaque année afin de déduire un comportement annuel distinct. En d'autres termes, nous développerons 10 modèles distincts, chacun correspondant à une année de 2012 à 2021. Notre approche d'apprentissage sera symbolique, reposant sur un arbre de décision, un générateur de règles d'induction/décision ou l'Analyse de Concepts Formels.

Cette méthodologie nous offre la possibilité d'interpréter et d'expliquer les modèles générés ultérieurement. Pour chaque année, nous obtiendrons un modèle d'apprentissage distinct. L'analyse de ces 10 modèles nous permettra de mieux comprendre l'évolution de la criminalité et des menaces telles que l'agitation, les émeutes, les attentats terroristes, et autres. Cette analyse consistera à comparer les arbres de décision ou les ensembles de règles générés, deux par deux. Il convient de noter que cette approche a fait ses preuves dans des travaux antérieurs (Meddouri et al., 2022).

Références

Da Silva, A. R. C., I. C. de Paula Júnior, T. L. C. da Silva, J. A. F. de Macêdo, et W. C. P. Silva (2020). Prediction of crime location in a brazilian city using regression techniques. In *Proceeding of IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI'20)*, pp. 331–336.

Meddouri, N., F. Rioult, et B. Cremilleux (2022). Apprentissage de comportements à partir de données temporelles hétérogènes. In *Proceeding of the EGC workshop Gestion et Analyse de données Spatiales et Temporelles with the 22^{ième} Journées Francophones en Extraction et Gestion des Connaissances (GAST/EGC'22)*.

Tolan, G., T. Abou-El-Enien, et M. Khorshid (2015). Hybrid classification algorithms for terrorism prediction in middle east and north africa. *International Journal of Emerging Trends Technology in Computer Science (IJETTCS) 4*, 23–29.

Summary

L'intelligence artificielle est un outils incontournable de nos jours pour analyser des données et apprendre des comportements. Une analyse spatio-temporelle des données de criminalité a suscité pas mal de travaux durant la dernière décennie. Malgré la disponibilité en libre accès des données de criminalité en France sur la période de 2011 jusqu'à 2021, aucun travail de recherche n'a traité ces données, ni considéré cette problématique. Dans ce papier, nous proposons un pré-traitement de ces données (de 2012 jusqu'à 2021) pour une analyse prédictive et un apprentissage interprétable des comportements criminels en vu d'une prévention des émeutes ou des attaques terroristes.

Écrêter la valeur cible ou filtrer les données en maintenance prévisionnelle : exemple de C-MAPSS

Résumé. Estimer la durée de vie restante, en anglais *Remaining Useful Life* (RUL), permet de suivre la santé d'un système pour limiter les coûts de sa maintenance, mais elle n'est réalisable que lorsque la dégradation du système a débuté. L'écrêtage de la valeur cible est souvent employé afin de limiter les impacts de la partie stationnaire des données, mais ce pré-traitement qui conserve les données impertinentes peut en partie limiter les performances du modèle. Nous proposons de filtrer les données d'entraînement du fonctionnement normal plutôt que d'écrêter la valeur cible, afin d'améliorer la performance du modèle. Nous évaluons l'efficacité de notre méthode sur le jeu de données C-MAPSS, souvent utilisé dans le domaine. La comparaison des résultats montre que notre approche possède une erreur plus faible, et cela encore plus selon la métrique ad-hoc à ce jeu de données.

1 Introduction

Des défaillances surviennent régulièrement dues à l'utilisation intensive et à l'usure des différentes machines. Celles-ci nécessitent ainsi de recourir à des actions de maintenance inopinées mettant à l'arrêt toute la production. La maintenance prévisionnelle consiste à prédire lorsqu'une panne risque d'arriver sur une machine, afin d'être capable de prévoir la maintenance associée et d'économiser les coûts d'une maintenance effectuée trop souvent (maintenance préventive) et d'une maintenance effectuée trop tard (maintenance corrective). Afin d'anticiper cette maintenance, on prédit une *Remaining Useful Life* (RUL), correspondant au temps avant la défaillance.

Lorsque l'on souhaite prédire le RUL, il n'est souvent possible de détecter l'apparition d'une défaillance que tardivement, car la grande majorité des données acquises correspondent à des données sans défaillance, et celle-ci n'est pas prédictible avant que la dégradation ait commencée. De nombreuses erreurs sont ainsi faites si l'on entraîne un modèle en utilisant des données avant la dégradation. On parle ainsi de "phase normale" ou "phase stationnaire" lorsque peu de variations sont présentes dans les données et que le phénomène de dégradation ou d'usure de la machine n'a pas encore débuté, et de "phase de dégradation" ou "phase dégradée" dès que celui-ci est perceptible.

Pour la suite de l'article, afin de visualiser des prédictions de RUL sur le jeu de données décrit dans la section suivante, on utilisera une forêt aléatoire avec les paramètres par défaut, à

savoir 100 arbres avec un minimum de 1 échantillon par feuille, entraînés jusqu'à ce que toutes les feuilles soient pures ou qu'elles contiennent moins de deux échantillons. La fenêtre choisie est de taille 15 afin de prendre en compte l'aspect temporel des données, on prendra donc en considération pour un individu les valeurs des paramètres des 15 pas de temps précédents.

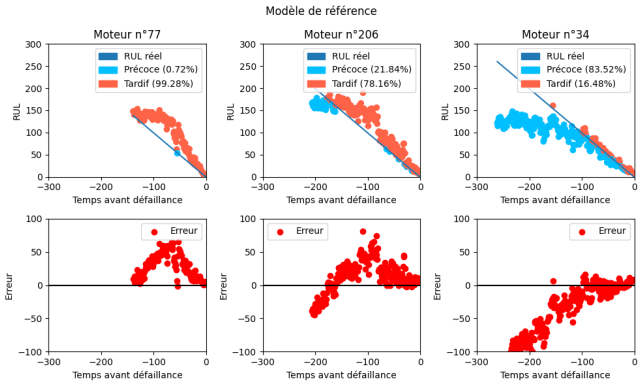


FIG. 1 – *Visualisation de prédictions sans pré-traitement sur 3 individus du jeu de données C-MAPSS*

La Figure 1, correspondant à des exemples de prédictions réalisées sur des individus du jeu de données C-MAPSS, permet de visualiser le phénomène d'impossibilité à prédire correctement. La première ligne de la figure correspond à des prédictions sur 3 individus, avec en rouge les prédictions trop élevées réalisées par le modèle, à savoir des prédictions tardives, et en bleu les prédictions trop faibles, correspondant à des prédictions précoces. La seconde ligne de la figure permet d'observer les erreurs réalisées par le modèle.

On peut constater sur les trois colonnes que si l'on entraîne un modèle sans prendre en compte cette spécificité, le modèle se trompe lourdement, atteint un plateau et commence à diminuer son erreur sur les prédictions. Le modèle ayant appris à prédire un RUL sur une quantité importante de données ne comportant aucune dégradation perceptible, beaucoup d'erreurs sont réalisées, et c'est seulement une fois avoir atteint cette phase de dégradation qu'il est capable de diminuer son erreur. Ce phénomène est d'autant plus perceptible lorsque la taille des séquences augmente. Sur la troisième colonne, on constate qu'une bonne partie des prédictions réalisées sont autour de 125 lorsque l'on se situe dans la phase normale, et que les erreurs commencent à diminuer beaucoup plus tard. Il est donc nécessaire de se focaliser uniquement sur la phase de dégradation et non sur la totalité des données si l'on cherche à avoir un modèle performant.

Dans le cas du jeu de données C-MAPSS, la phase de dégradation est souvent fixée à 125 ou 130, car il s'agit de la valeur utilisée par Heimes (2008) dans son approche, largement citée dans la littérature. Nous considérerons donc un seuil à 125 dans la suite de l'article.

Celui-ci étudie l'influence de la préparation des données sur l'apprentissage d'un modèle de prédiction de la durée de vie restante (RUL). Les contributions de cet article sont les suivantes : 1) Nous montrons que les méthodes utilisées dans l'état de l'art pour prendre en compte l'état normal / défaillant contribuent à déformer le modèle, et ainsi à dégrader les performances de prédiction du modèle. 2) Nous proposons une méthode pour y répondre consistant à réaliser

l'apprentissage uniquement sur les données défaillantes, que nous validons expérimentalement sur le jeu de données C-MAPSS. La suite de l'article est décomposée comme suit : la section 2 décrit le jeu de données utilisé, la section 3 détaille l'usage de l'écrêtage dans l'état de l'art, la section 4 introduit notre approche et sa plus-value par rapport à l'état de l'art, la section 5 présente les résultats des différents modèles et en quoi notre approche est plus adaptée et performante, et la section 6 conclut et propose quelques perspectives.

2 C-MAPSS

C-MAPSS est un jeu de données d'essai (*benchmark*) typique du domaine de la maintenance prévisionnelle (Ramasso et Saxena, 2014). Développé par le *Prognostics Center of Excellence* de la NASA et décrit dans Saxena et al. (2008), il s'intéresse à des simulations de dégradation de turboréacteurs.

Sous-jeux de données	FD001	FD002	FD003	FD004
Nombre de séquences d'entraînement	100	260	100	249
Nombre de séquences de test	100	259	100	248
Conditions de vol	1	6	1	6
Conditions de défaillance	1	1	2	2
Nombre de valeurs de capteurs		21		

TAB. 1 – *Caractéristiques du jeu de données C-MAPSS*

Il est composé de 4 sous-jeux de données comme décrit dans le Tableau 1, à savoir FD001, FD002, FD003 et FD004. Chacun de ces sous-jeux de données est divisé en séquences d'entraînement et de test, selon une répartition 50% / 50% ; les sous-jeux de données FD001 et FD003 sont composés de 200 séquences, et les jeux de données FD002 et FD004 sont constitués respectivement de 519 et 497 séquences.

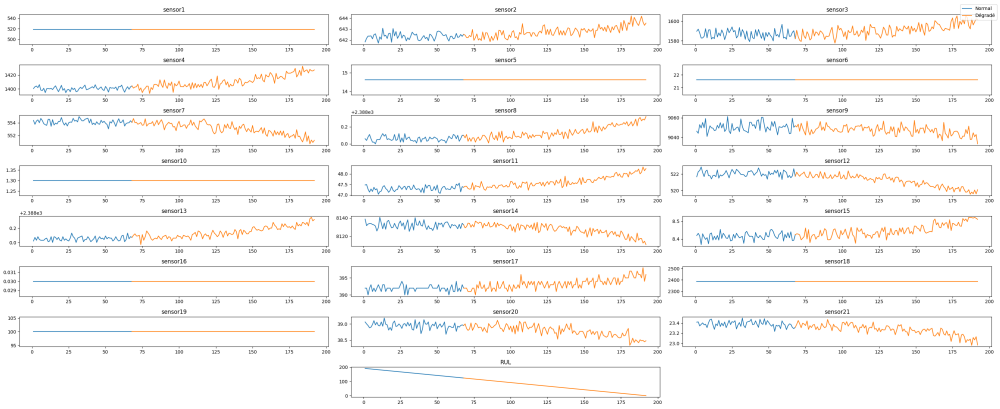


FIG. 2 – *Visualisation des valeurs des paramètres pour le premier moteur de FD001*

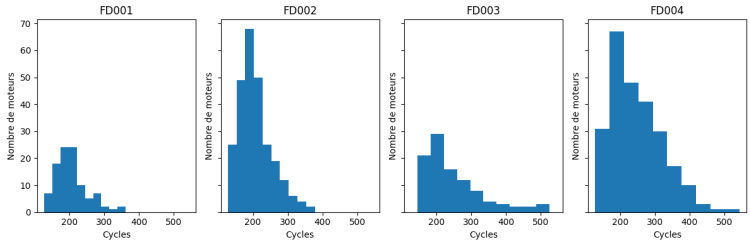


FIG. 3 – *Nombre de cycles maximal des moteurs des différents sous-jeux de données*

Chaque moteur est décrit par une séquence de mesures de l'état du moteur (cycles), jusqu'à ce que le moteur soit défaillant. La répartition des longueurs des séquences d'entraînement pour chacun jeu de données est décrite en Figure 3. Les jeux de données sont majoritairement composés de moteurs ayant une durée de vie moyenne autour de 200 cycles, avec les sous-jeux de données FD001 et FD002 ayant une répartition similaire de taille de séquences et un nombre de cycle maximal autour de 350 cycles, là où les sous-jeux de données FD003 et FD004 ont une répartition plus étalée avec un nombre de cycles maximal de 520 cycles.

Les moteurs sont également décrits par un numéro, les données des 21 capteurs, les conditions de vol, et le nombre de cycles depuis lequel il est fonctionnel. La Figure 2 permet de visualiser l'évolution des différents paramètres pour le moteur n°1 du sous-jeu de données FD001. En bleu sont représentés les valeurs des paramètres lors du fonctionnement normal de la machine, et en orange les valeurs lors du fonctionnement dégradé.

Ce jeu de données est largement utilisé dans le domaine de la maintenance prévisionnelle car il contient un grand nombre de machines différentes ayant subi des dégradations, et car un nombre important de paramètres permettent de caractériser l'évolution de l'état du système.

3 Travaux connexes

Comme précisé dans la section 1, sans utiliser de pré-traitement, il n'est pas possible pour le modèle de prédire un RUL de manière efficace car beaucoup de données ne comportant pas de dégradation sont utilisées par le modèle. La méthode utilisée habituellement pour réduire l'impact des données collectées en fonctionnement normal consiste à écrêter le RUL (Heimes, 2008) afin qu'il ne dépasse pas un seuil. Plus les séquences sont longues, plus le RUL croît. Écrêter permet ainsi de réduire la plage de valeurs du RUL en transformant les valeurs supérieures au seuil tel que défini dans l'équation 1.

$$RUL_{ecretage} = \begin{cases} \text{seuil}, & RUL > \text{seuil} \\ RUL, & \text{sinon} \end{cases} \tag{1}$$

De nombreuses approches utilisent l'écrêtage dans la littérature. On retrouve ainsi l'usage de cette approche, notamment avec un seuil à 125, de manière très commune comme dans Benker et al. (2021), Li et al. (2018) et Wang et al. (2018).

Certaines approches utilisent un seuil d'écrêtage différent pour chaque sous-jeu de données, comme Sateesh Babu et al. (2016), sans toutefois préciser lesquels. Listou Ellefsen et al. (2019) proposent cependant une approche en utilisant différents seuils entre 115 et 140.

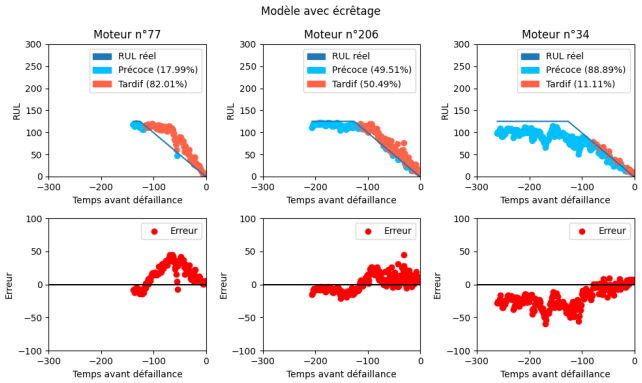


FIG. 4 – *Visualisation de prédictions avec écrêtage sur 3 individus*

Peu d'approches n'utilisent pas l'écrêtage. Elles cherchent au mieux à définir des seuils de manière plus automatisée. Yu et al. (2020) estiment par exemple qu'il est difficile d'estimer une valeur universelle applicable pour tous les individus et que cela apporte beaucoup d'erreur pour ceux qui se dégradent rapidement. Lan et al. (2018) proposent une méthode basée sur la distance euclidienne pour identifier le seuil d'écrêtage à utiliser pour améliorer les prédictions.

L'utilisation de cette méthode suppose que l'on connaisse le seuil à partir duquel écrêter d'après les données, ce qui n'est pas toujours le cas. Cependant, que le seuil soit fixé de façon arbitraire ou détecté par apprentissage automatique, la question est savoir de comment l'utiliser et la plupart des approches existantes écrêtent les valeurs cibles supérieures au seuil retenu.

La Figure 4 montre un exemple de prédiction avec écrêtage sur les mêmes moteurs que la Figure 1. Nous observons sur les colonnes de celle-ci que modifier la valeur de la variable cible, sur les données d'entraînement et sur les données de test, diminue l'erreur sur la phase de fonctionnement normal, ce qui était l'effet recherché, mais déforme les prédictions sur la phase dégradée également. Le modèle, pour minimiser son erreur, va prédire très souvent des valeurs autour du seuil d'écrêtage, ici 125. Plus la taille des séquences augmente, et plus la dégradation des prédictions s'accentue. La troisième colonne montre l'impact ce phénomène, avec des prédictions qui fluctuent énormément avant la phase de dégradation, alors même que le RUL reste stable. Ainsi, même si le modèle effectue moins d'erreur en moyenne car la valeur du RUL a été modifiée, les performances n'en sont pas meilleures pour autant. Les sections suivantes présentent plus en détails notre approche et la comparaison que nous faisons avec l'écrêtage et la baseline.

4 Notre approche

Dans la section 1, nous avons vu qu'utiliser les données de la phase normale conduit à plus d'erreurs, car il n'est pas possible de prédire le RUL dans cette phase, et le modèle ainsi

entraîné commet plus d'erreurs aussi sur les données de la phase dégradée. Nous avons vu dans le section 3 que l'écrêtage réduit les erreurs sur la phase normale, mais conduit à déformer le modèle y compris sur la phase dégradée. Nous supposons également que l'on sait distinguer les données de la phase normale de la phase dégradée.

Algorithme 1 : Filtrage des données

Entrées : Données non traitées
Sorties : Données filtrées
créer une liste de données vide ;
tant que *toutes les données ne sont pas filtrées* **faire**
  récupérer la donnée non-traitée suivante ;
  si *donnée dans phase dégradée* **alors**
    ajouter la donnée dans la liste ;
  fin
fin

Nous proposons ainsi une approche, appelée filtrage. Elle consiste à considérer uniquement les données de la phase dégradée et ce sans changer les valeurs de la variable cible. Pour C-MAPSS, la valeur de 125 pas de temps est choisie étant donné qu'il s'agit du seuil le plus utilisé dans la littérature comme précisé dans la section 1.

Le fait de filtrer les données permet de s'intéresser uniquement aux informations correspondant à une dégradation, sans prendre en compte les données de fonctionnement normal qui empêcheraient le modèle d'extraire les caractéristiques présentes dans les données. Le fonctionnement du filtrage est résumé dans l'Algorithme 1.

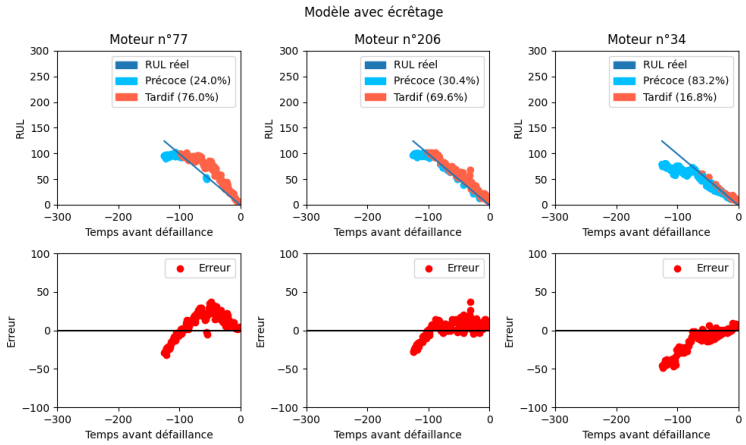


FIG. 5 – *Visualisation de prédictions avec filtrage sur 3 individus*

La Figure 5 permet de visualiser des exemples de prédictions réalisées avec les données filtrées. On constate que les prédictions sont plus resserrées et plus proches de la vérité terrain. Le modèle ayant été entraîné seulement sur des données comportant des dégradations, il n'a

pas été brouillé par les données du fonctionnement normal. Cela permet au modèle d'être plus performant sur la phase de dégradation car il n'a été entraîné que sur ces données qui comportent des variations perceptibles, mais également d'éviter d'ajouter de l'erreur en prédisant un RUL approximatif pour la phase stationnaire.

5 Comparaisons expérimentales

Nous comparons l'algorithme défini dans la section 1 entraîné à partir de données préparées de trois manières différentes :

— sans traitement particulier afin d'avoir une *baseline*,
— avec écrêtage de la valeur cible à 125,
— sans écrêtage, mais en filtrant les données pour avoir les 125 derniers pas de temps de chaque moteur.

Étant donné que les jeux de test inclus dans C-MAPSS ne contiennent que le RUL à prédire à un instant précis et pas toute l'évolution de la dégradation, nous avons choisi de nous intéresser qu'aux jeux d'entraînement. Le pré-traitement effectué est très simple pour éviter au maximum les biais du modèle. Nous avons sélectionné 15 paramètres parmi les 27 (21 valeurs de capteurs + 6 conditions de vol), car ce sont elles qui apportent le plus d'information, la distribution des autres paramètres ayant un écart-type quasiment égal ou égal à 0. Nous avons ensuite divisé les jeux utilisés en jeux d'entraînement et jeux de test avec une répartition 80% / 20%.

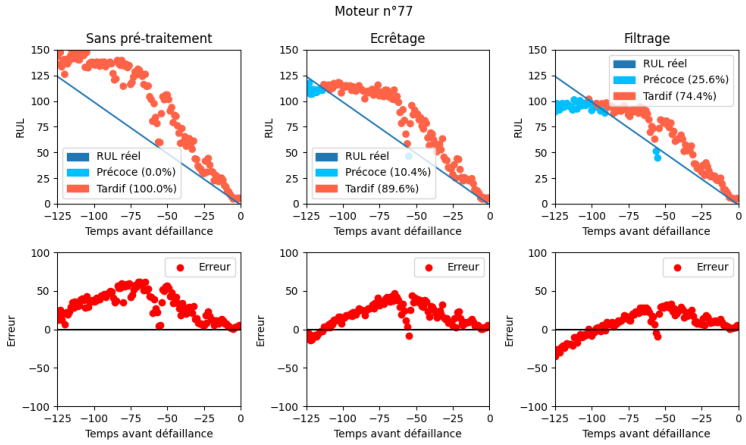


FIG. 6 – *Visualisation des prédictions sur un moteur du sous-jeu de données FD001*

Les Figures 6, 7, 8 et 9 permettent de visualiser l'impact des différentes méthodes de préparation des données sur des séquences, respectivement de petite, moyenne et grande taille, représentant différents scénarios dans les données. Il s'agit également des prédictions visibles sur les Figures 1, 4 et 5.

Écrêter la valeur cible ou filtrer les données en maintenance prévisionnelle

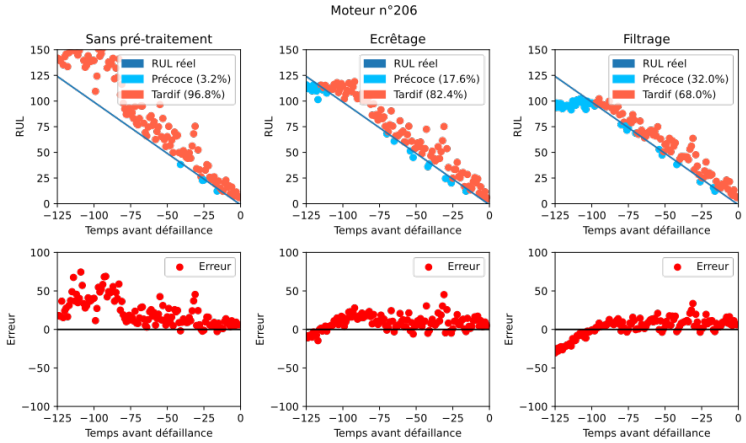


FIG. 7 – *Visualisation des prédictions sur un moteur du sous-jeu de données FD002*

Sur la colonne de droite de la Figure 6, on constate que le fait de filtrer les données a permis de recentrer les prédictions vers le RUL réel, en réduisant la sur-estimation faite par le modèle, avec 27% pour le modèle sans pré-traitement et de 9% pour celui avec écrêtage.

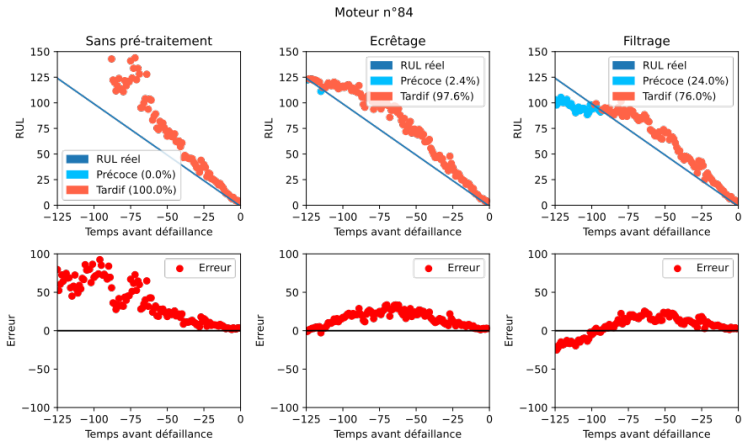


FIG. 8 – *Visualisation des prédictions sur un moteur du sous-jeu de données FD003*

Dans le cas de séquences de taille moyenne, telles que les Figures 7 et 8, le filtrage permet également de réduire légèrement l'erreur une fois la phase de dégradation atteinte, tout en supprimant les sous-estimations faites avant cette phase. Dans le cas d'une grande séquence, comme sur la Figure 8, on constate que les modèles sans pré-traitement et avec écrêtage font des prédictions précoces sur un grand nombre de pas de temps, notamment pour le second étant donné que la valeur cible d'une grande partie de la séquence a été réduite à 125. Les modèles

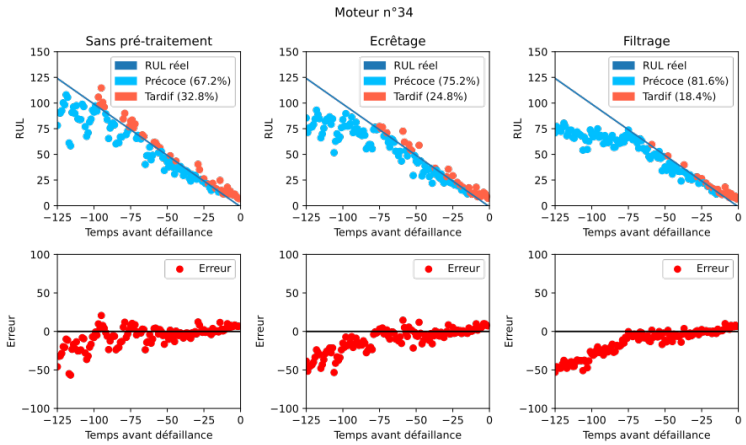


FIG. 9 – *Visualisation des prédictions sur un moteur du sous-jeu de données FD004*

n'étant pas capables de prédire une valeur avant d'avoir atteint la phase de dégradation, ils prédisent un RUL autour de 125 car il s'agit de la valeur la plus commune dans les données.

Pour le modèle avec filtrage, le fait de se concentrer uniquement sur les données de dégradation a permis de réduire l'erreur totale en évitant de prédire une valeur autour de 125 avant le début de la défaillance. Ainsi, le modèle évite d'ajouter de l'erreur en prédisant des valeurs qu'il ne peut pas trouver, mais il est également plus performant sur la phase dégradée.

Nous comparons également les approches globalement en mesurant la performance des modèles à l'aide des métriques classiquement utilisées sur C-MAPSS et pour les problématiques de régression, à savoir l'erreur absolue moyenne (MAE), la racine carrée de l'erreur quadratique moyenne (RMSE), le coefficient de détermination (R2), ainsi que la métrique d'évaluation proposée avec C-MAPSS, appelée RUL Score.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2a}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2b}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2c}$$

$$S = \begin{cases} \sum_{i=1}^{n} e^{(\frac{y_i - \hat{y}_i}{10})} - 1, & y_i - \hat{y}_i < 0 \\ \sum_{i=1}^{n} e^{-(\frac{y_i - \hat{y}_i}{13})} - 1, & y_i - \hat{y}_i \geq 0 \end{cases} \tag{2d}$$

Les deux premières métriques permettent de connaître de manière absolue et relative l'erreur totale réalisée par le modèle, nous permettant d'avoir une idée plus précise de la performance absolue des modèles. Le coefficient de détermination permet lui de mesurer avec un

score entre 0 et 1 la qualité de la prédiction réalisée et sa proximité avec la réalité, en comparant le gain de performance par rapport à une prédiction qui serait faite en estimant uniquement la moyenne. Enfin, le RUL Score est une métrique ad-hoc à C-MAPSS, visant à pénaliser plus fortement les prédictions tardives que précoces, car dans une utilisation réelle, une prédiction tardive est beaucoup plus coûteuse qu'une prédiction précoce dû à la nécessité de prévoir la maintenance suffisamment tôt pour pouvoir l'effectuer convenablement. Cette dernière vise à pénaliser plus fortement des prédictions tardives que des prédictions précoces, ce que l'on cherche à faire dans un cadre de maintenance prévisionnelle où l'arrêt inopiné d'un système est très coûteux. Les métriques sont définies dans l'équation 2.

FD001			
	Baseline	Ecrêtage	Filtrage
MAE	19.27	12.52	**11.86**
RMSE	26.86	17.26	**15.75**
R^2	0.44	0.77	**0.81**
RUL Sc.	157489	19373	**9656**

FD002		
Baseline	Ecrêtage	Filtrage
26.57	16.45	**14.31**
33.69	21.19	**17.87**
0.13	0.65	**0.75**
$1.27.10^6$	109625	**36685**

FD003			
	Baseline	Ecrêtage	Filtrage
MAE	23.16	12.67	**11.82**
RMSE	37.91	18.20	**15.83**
R^2	-0.10	0.75	**0.81**
RUL Sc.	$2.47.10^8$	31247	**10783**

FD004		
Baseline	Ecrêtage	Filtrage
28.85	16.45	**14.31**
40.61	21.19	**17.87**
-0.27	0.65	**0.75**
$6.69.10^7$	130735	**55779**

TAB. 2 – *Performance des modèles sur les différents jeux de données*

avec y_i la valeur prédite de i, $\hat{y}_i$ la valeur théorique de i et $\bar{y}$ la moyenne des valeurs théoriques. Pour chaque sous-jeu de données, on mesure la performance des différents modèles. Étant donné l'hypothèse précisée dans la section 4 que les 125 derniers pas de temps correspondent à la défaillance, seuls ceux-ci sont pris en compte pour comparer les résultats présentés dans le Tableau 2.

Sous-jeu de données	Baseline	Écrêtage	Filtrage
FD001	25% / 75%	37% / 63%	**48% / 52%**
FD002	14% / 86%	26% / 74%	**40% / 60%**
FD003	27% / 73%	37% / 63%	**51% / 49%**
FD004	23% / 77%	34% / 66%	**48% / 52%**

TAB. 3 – *Comparaison du pourcentage de prédictions précoces / tardives des différents modèles sur les 125 derniers pas de temps*

On constate que les modèles avec écrêtage et avec filtrage des données ont de meilleures performances que la baseline, mais le modèle avec filtrage reste plus performant, notamment avec le RUL Score qui est 2 à 3 fois plus faible. La visualisation de la répartition des prédictions

entre précoces et tardives, comme présentée dans le Tableau 3, permet également d'avoir une visualisation de la manière dont les modèles performent.

On peut constater que le fait d'écrêter engendre une forte sous-estimation du RUL pour diminuer l'erreur sur la totalité des données, et une forte sur-estimation sur les derniers pas de temps. La majeure partie des valeurs cibles étant à 125, le modèle fait beaucoup de prédictions précoces. Ce phénomène s'accentue sur le modèle de base, où le fait d'être entraîné sur la totalité des données empêche le modèle de se calibrer correctement.

Le filtrage des données conduit à un ratio plus proche de 50% / 50%, montrant que notre modèle a une estimation plus proche de la réalité en calibrant sa prédiction autour de la valeur cible.

6 Conclusion

L'approche développée dans cette étude montre comment améliorer la prédiction du Remaining Useful Life d'un système et le valide expérimentalement sur le jeux de données de référence C-MAPSS. Les résultats montrent que le filtrage des données permet d'améliorer significativement les performances du modèle par rapport au modèle de base. Pour pouvoir faire une prédiction efficace du RUL, il est ainsi nécessaire de ne pas prendre en compte les données avant la phase de dégradation. Notre approche part d'une hypothèse forte, à savoir que l'utilisateur a connaissance de l'instant séparant la phase normale de la phase de dégradation. Or en pratique, celui-ci n'est pas nécessairement connu. Dans nos prochains travaux, nous chercherons à développer une approche non-supervisée permettant de détecter l'apparition de la phase de dérive du système. Enfin, les prédictions étant actuellement indépendantes, la prise en compte de cet aspect à travers l'apprentissage par un modèle récurrent permettrait de lisser le modèle, et ainsi d'améliorer les performances et la fiabilité des prédictions.

Références

Benker, M., A. Bliznyuk, et M. F. Zaeh (2021). A Gaussian Process Based Method for Data-Efficient Remaining Useful Life Estimation. *IEEE Access 9*, 137470–137482. Conference Name : IEEE Access.

Heimes, F. O. (2008). Recurrent neural networks for remaining useful life estimation. In *2008 International Conference on Prognostics and Health Management*, pp. 1–6.

Lan, G., Q. Li, et N. Cheng (2018). Remaining Useful Life Estimation of Turbofan Engine Using LSTM Neural Networks. In *2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC)*, pp. 1–5.

Li, X., Q. Ding, et J.-Q. Sun (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety 172*, 1–11.

Listou Ellefsen, A., E. Bjørlykhaug, V. Æsøy, S. Ushakov, et H. Zhang (2019). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. *Reliability Engineering & System Safety 183*, 240–251.

Ramasso, E. et A. Saxena (2014). Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets. *International Journal of Prognostics and Health Management 5*(2). Number : 2.

Sateesh Babu, G., P. Zhao, et X.-L. Li (2016). Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life. In S. B. Navathe, W. Wu, S. Shekhar, X. Du, X. S. Wang, et H. Xiong (Eds.), *Database Systems for Advanced Applications*, Lecture Notes in Computer Science, Cham, pp. 214–228. Springer International Publishing.

Saxena, A., K. Goebel, D. Simon, et N. Eklund (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management*, pp. 1–9.

Wang, J., G. Wen, S. Yang, et Y. Liu (2018). Remaining Useful Life Estimation in Prognostics Using Deep Bidirectional LSTM Neural Network. In *2018 Prognostics and System Health Management Conference (PHM-Chongqing)*, pp. 1037–1042. ISSN : 2166-5656.

Yu, W., I. Y. Kim, et C. Mechefske (2020). An improved similarity-based prognostic algorithm for RUL estimation using an RNN autoencoder scheme. *Reliability Engineering & System Safety 199*, 106926.

Summary

Estimating the Remaining Useful Life (RUL) allows to monitor the state of health of a machine in order to reduce the maintenance costs, but it can be achieved only when the degradation phenomenon has started. Clipping the target value is usually employed to limit the impact of the normal part of data, but limit the performance of the model as it keeps the irrelevant data. Our approach aims to increase the performance of the model by filtering this part of the train data instead of clipping the target value. We evaluate our method on the benchmark dataset C-MAPSS. Comparing the results shows that our approach achieves a lower error, and even more with the ad-hoc metric provided with the dataset.

Méthode de gestion et d'analyse de l'information spatiale et temporelle maritime

Anne-Clémence DUVERGER*, Cyril RAY*

*Institut de Recherche de l'École Navale
École Navale - CC 600
29240 Brest Cedex 9, FRANCE
prénom.nom@ecole-navale.fr

Résumé. Le suivi, l'analyse et la compréhension des mobilités maritimes par essence libre de circulation est un élément majeur de nos sociétés modernes tant elles témoignent des enjeux commerciaux, sociétaux ou géopolitiques qui existent entre les pays. Les volumes de données décrivant ces mobilités maritimes sont croissants, fluctuant et l'extraction et la gestion des connaissances au profit, par exemple, du suivi des pêches, des échanges commerciaux ou des actes illicites, etc. nécessitent une démarche d'étude complémentaire entre ingénierie de l'information géographique, intelligence artificielle et sciences humaines. Dans ce contexte, notre objectif est de développer des capacités pluridisciplinaires d'analyse et d'interprétabilité des mobilités maritimes (quels sont les navires, quelles sont leurs caractéristiques, etc.), d'analyse de leurs comportements maritimes (où se déplacent-ils, à quelle vitesse, etc.) et d'analyse de leurs situations maritimes (dans quel contexte se déplacent-ils, pourquoi, quel est le contexte international, etc.). Dans cet article, nous proposons une méthodologie d'analyse relevant de l'intelligence géospatiale et présentons l'architecture de gestion des données construites. Les travaux sont illustrés par une étude préliminaire des enjeux numériques maritimes de la zone Indo-pacifique.

Introduction

Nous avons assisté à une territorialisation continue de l'espace maritime (Miossec, 2014). En effet, les trajets les navires forment aujourd'hui de véritables « *autoroutes de la mer* » (Royer, 2012). Ces routes maritimes relient principalement l'Europe, l'Asie orientale et l'Amérique du Nord, néanmoins, les routes nord-sud ou sud-sud ne cessent de progresser. Le transport maritime assure aujourd'hui plus de 80 % des échanges mondiaux (Fremont et Fremont-Vanacore, 2014). Les profondeurs des océans sont aussi soumises à ce phénomène d'expansion des espaces anthropisés : l'océan est devenu un véritable « *front-pionnier* » (Miossec, 2014) (ZEE [1] en litige, exploitation de ressources naturelles, pose de câbles sous-marins...), engendrant ainsi de nouvelles conflictualités qui s'accompagnent d'un besoin accru de gestion de l'information maritime (qui est par nature spatiale et temporelle).

1. Zone économique exclusive

Pourtant, l'intelligence géospatiale [2] s'est jusqu'à présent peu saisie des problématiques liées à l'environnent maritime. Pour le géographe Philippe Boulanger, cette dernière se définit « *comme un processus de fusion de données géolocalisées et géo-référencées à partir d'une diversité de sources qu'elles soient militaires ou civiles* » (Boulanger, 2019a). De plus, elle est une discipline globale, qui permet la synthèse de différents champs de compétences, tant dans le domaine des sciences humaines : géopolitique, économie, sociologie, histoire, géographie, anthropologie, que dans celui des technologies (Morel et Boulanger, 2016).

Le CICDE (2021) [3] précise que l'intelligence géospatiale se caractérise par l'utilisation de données spatio-temporelles : « *données (...) géolocalisées, géoréférencées et préférentiellement horodatées (...)* » et est basée sur des « *procédés de calcul complexes* » en vue de procéder à des analyses spatiales. En effet, avec l'avènement d'un volume toujours plus massif de données (*Big Data* en anglais), les algorithmes deviennent incontournables pour le traitement de ces dernières. L'analyse spatiale maritime n'échappe pas à ce constat.

En effet, si les données spatio-temporelles des navires, notamment les données du système AIS [4], sont précieuses pour comprendre les activités maritimes, celles ci sont croissantes et volumineuses. L'agence européenne pour la sécurité maritime (EMSA), par exemple, annonce traiter environ 50 million de message de position par jour. Cela demande d'optimiser la mise en place de procédés adaptés à l'ensemble du processus d'exploitation de ces données spatio-temporelles : de leur stockage jusqu'à leur représentation en passant l'examen critique de leur apport à l'analyse des situations maritimes.

Dans nos travaux nous cherchons à poser les bases d'une méthodologie générique relevant de l'intelligence géospatiale maritime. Dans ce contexte, nous reviendrons premièrement sur la genèse de l'intelligence géospatiale maritime tout en participant à sa formalisation (section 1), puis nous décrirons la représentation des données spatio-temporelles maritimes *via* des graphes et des hexagones (section 2). La section 3 présente les mécanismes de traitement et de stockage des données ainsi que quelques algorithmes réalisés. Enfin une étude de cas relevant de l'intelligence géospatiale sera décrite (l'analyse de la vulnérabilité des câbles sous-marins en Indo-Pacifique) (section 4).

1 L'intelligence géospatiale et les données spatio-temporelles maritimes

1.1 L'émergence d'une intelligence géospatiale maritime

L'intelligence géospatiale est le fruit d'un retour, depuis les années 1990, d'une géographie militaire intégrant tous les composants géographiques et ce sous l'effet de l'importance donnée au renseignement et de l'émergence des « NTIC » [5] (Boulanger, 2006). En effet, l'information géographique repose aujourd'hui sur des principes nouveaux, liés à l'emploi des hautes technologies, à l'immédiateté des informations (Boulanger, 2011).

« *La Géo intelligence traite fondamentalement de théâtres terrestres* » (Prazuck, 2022) car elle a pris son essor durant les guerres du Koweït (1990-1991), de Bosnie (1991-1995), du

2. Dans le sens anglo-saxon du terme intelligence ; i.e. renseignement.
3. Centre interarmées de concepts, de doctrines et d'expérimentations
4. Automatic Identification System
5. Nouvelles technologies de l'information et de la communication

Kosovo (1999), d'Irak (2003) et d'Afghanistan (2001) (Boulanger, 2020). Ce n'est que plus récemment qu'elle a été appliquée aux opérations maritimes. Dans ce domaine, l'Inde est pionnière ; en 2011, elle crée le *National Command Control Communication and Intelligence* [6]. De plus, depuis 2007 des conférences dédiées à l'intelligence géospatiale sont organisées (sur le modèle de celles de l'USGIF [7]) ; dans le cadre du partenariat stratégique entre ce pays et les États-Unis (Boulanger, 2019b). La dernière, intitulée « *Indo-Pacifique Geointelligence 2023* » organisée en juin 2023 à New Delhi, réunissait des responsables gouvernementaux, des militaires, des experts des domaines du spatial, géospatial, de l'industrie et des universitaires.

En effet, l'intelligence géospatiale a pénétré le monde académique, élargissant de ses champs de recherches : santé (Gehlen et al., 2019), protection de l'environnement (Abu Sari et al., 2018), gestion de crise (Markogiannaki et al., 2020), aménagement du territoire (Parambil et al., 2021). En France, les études relevant de l'intelligence géospatiale appliquée sont encore rares et dans le domaine maritime, seuls quelques travaux s'en approchent : César Ducruet (2013, 2015, 2016).

1.2 L'intelligence géospatiale maritime, essaie de définition

En s'appuyant sur les travaux du CICDE, nous avons identifié les principaux composants caractérisant l'intelligence géospatiale maritime : elle exploite du renseignement multisources ; c'est (1) une approche globale qui permet (2) la contextualisation des données issues de toutes sources ; (3) les données sont exploitées dans des SIG [8] et *via* (4) des procédés d'automatisation de calculs ; elle permet (5) l'analyse prédictive et l'anticipation ; elle est essentielle (6) à la planification ; elle est (7) assimilable à un système d'information, car elle permet de collecter, entreposer, traiter et distribuer de l'information ; enfin, (8) elle structure des flux d'informations provenant d'acteurs privés ou institutionnels.

Rajoutons que l'intelligence géospatiale maritime doit considérer les quatre dimensions de ce milieu : la surface, la colonne d'eau, le plancher océanique et l'espace, en raison de son degré d'anisotropie [9] moins important que celui de l'espace terrestre ; il existe moins de structures de surface sur lesquelles se baser pour caractériser une situation.

Ainsi, l'intelligence géospatiale maritime s'appuie sur différents capteurs générant des informations issues des quatre dimensions de ce milieu (cf. figure 1). Sur les mers et océans (surface), les navires sont dotés de systèmes de localisation : Automatic Identification System (AIS) (destiné à tous les navires), Vessel monitoring system (VMS) qui est système de surveillance des navires de pêche ou encore Voyage Data Recorder (VDR) qui est un enregistreur de données du voyage. En surface, on peut également collecter des données issues du renseignement humain, des caméras des ports et des RADAR [10]. D'autres types de données sont disponibles pour l'environnement maritime immergé : données des bathysondes, SONAR [11],

6. Ce projet de sécurité maritime, lancé après les attentats de Mumbai (2008), s'appuie sur un système de surveillance qui comprend des radar, caméras, avions, satellites et le système d'identification automatique des navires (AIS). Les données issues de ces capteurs sont collectées par le centre de gestion et d'analyse de l'information de la marine indienne (IMAC).

7. United States Geospatial Intelligence Foundation

8. Système d'information géographique

9. L'anisotropie caractérise un espace orienté, qui s'ordonne selon des axes, qui obéit à des polarisations (Tabarly, 2005) ; il n'existe pas d'espace géographique isotrope.

10. Radio detection and ranging, détection et estimation de la distance par ondes radio

11. Sound Navigation and Ranging, système de navigation et de télémétrie par écho sonore

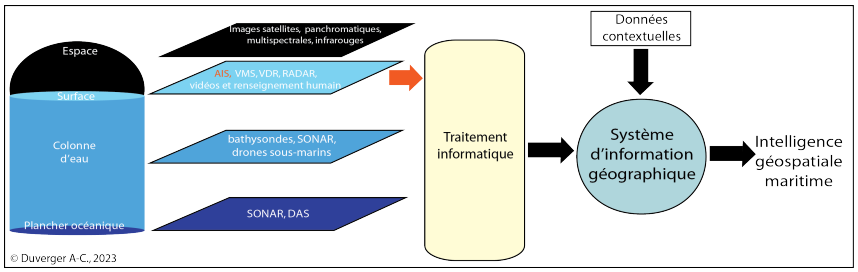


FIG. 1 – L'intelligence géospatiale maritime : milieu physique, données et méthode

des drones sous-marins et DAS [12]. Enfin, les satellites se déplaçant au-dessus de l'espace maritime sont une source de données importante.

Au regard des sources d'informations disponibles et en se basant sur les caractéristiques physiques du milieu maritime, l'intelligence géospatiale maritime peut se définir comme une méthode d'analyse qui permet d'acquérir des connaissances par la fusion de toutes ou une partie des données acquises dans les quatre dimensions du milieu maritime. Certaines d'entre elles sont particulières au milieu maritime, notamment les données AIS.

Ces dernières représentent un volume d'informations important. Ainsi, elles nécessitent des systèmes de stockage adaptés et doivent être traitées par des algorithmes pour faciliter leur représentation. Il est ensuite possible de les transférer dans un SIG, afin de les spatialiser et de procéder à la production de cartes. D'autres types de représentations (graphiques et schémas) permettent d'enrichir la compréhension de la dimension spatiale et temporelle de l'information.

Rajoutons que l'intégration de données contextuelles relatives au milieu maritime permet d'enrichir l'analyse tout facilitant l'interprétation des données de localisation de navires, car celles-ci sont partielles et pas toujours fiables : erreurs, falsification, usurpation des messages (Iphar et al., 2015).

2 Représentation des données spatio-temporelles maritimes

2.1 Des données massives

Notre méthodologie d'analyse pour l'intelligence géospatiale a été implémentée et testée avec un jeu de données contenant les localisations de navires dans le monde durant l'année 2019. Ces données sont obtenues grâce au système AIS (*Automatic Identification System*). L'AIS est un dispositif destiné à la prévention des collisions en mer mais également permettant un suivi en temps réel navires [13]. Les navires qui en sont équipés (environ 400 000 à l'échelle mondiale) transmettent régulièrement leur position (quelques secondes) ainsi que leur identité et leurs informations de voyage (quelques minutes) [14].

12. Distributed Acoustic Sensing, système de détection acoustique distribuée.
13. Accessible par exemple via le portail Web Marine Traffic : https://www.marinetraffic.com.
14. Numéro MMSI (Marine Mobile Service Identity), numéro IMO (identifiant de l'Organisation maritime internationale), nom, type de navire, dimensions, tirant d'eau, position (latitude et longitude), vitesse sur le fond, cap, cap sur le sol, rayon de giration, dangerosité de la cargaison, port de destination, temps estimé d'arrivée).

Le jeu de données intègre deux versions : les messages AIS d'origines (environ 1,7 Go par jour) et les données AIS prétraitées, sous-échantillonnées à 5 minutes mais agrégées avec les données de la *Lloyds* (environ 6,5 Go par jour). Ce deuxième format constitue notre base de travail principale car elle apporte 37 champs complémentaires aux données AIS ; e.g. propriétaire, opérateur, année de construction du navire, etc. Le volume de données rend difficile la fouille des données historiques et l'extraction d'informations pertinentes par un utilisateur. Nous avons mis en oeuvre deux approches d'abstraction des données AIS historiques ; une indexation hexagonale et un graphe hiérarchiques.

2.2 Représentations hexagonales des données maritimes

Nous avons basé une partie des algorithmes et cartographies sur un système d'indexation géospatiale hiérarchique hexagonal (selon le modèle Uber H3) qui rend la solution originale à plus d'un titre par rapport aux travaux existants (Maslek Elayam et al., 2022). Les hexagones sont utilisés dans plusieurs traitement en remplacement des coordonnées géographiques précises des navires ou pour construire une connaissance sur le temps long (statistique) de l'espace maritime.

Ainsi, l'espace maritime est divisé en cellules hexagonales d'une résolution donnée (la résolution est paramétrable). L'hexagone est une structure de données adaptée car permettant de stocker des statistiques maritimes historiques telles que le nombre de navires ayant traversé chaque hexagone pendant une certaine période (densité de l'hexagone) et leur vitesse minimale, maximale et moyenne. Ces statistiques et bien d'autres liées à la trajectoire fournissent aux algorithmes des informations utiles pour vérifier les comportements réguliers et anormaux.

La précision et la pertinence des statistiques calculées dépendent de plusieurs facteurs tels que la taille des données sur zone, la résolution de l'hexagone et l'intervalle de temps (microscopique : par heure, mésoscopique : par jour ou macroscopique : par semaine ou par mois).

Au-delà des bénéfices algorithmiques de cette approche basée sur les hexagones, elle a démontré un avantage certain dans la réduction des volumes de données et dans l'accroissement des performances computationnelles.

2.3 Représentations par modèle de graphe

Afin d'identifier des récurrences et certaines anomalies dans les trajets des navires nous avons opté pour la définition d'un modèle de graphe (Maslek Elayam et al., 2021). Dans ce dernier, les nœud représentent les origines et destinations et les points saillants d'une trajectoire de navire. Les arcs sont les flux entre ces nœuds. Le modèle de graphe conçu est hiérarchique, ainsi il peut être construit à partir de données AIS historiques (cf. Figure 2). Des fonctions d'agrégation permettent d'abstraire le graphe spatio-temporel à différentes échelles spatiales et temporelles. La dimension spatiale (i.e. les positions des navires) est réalisée sous forme d'une indexation sémantique par grille régulière et hiérarchique (i.e. avec les hexagones H3).

L'objectif de cette structure de données est double. Premièrement cela permet l'exploration et la compréhension du réseau maritime ainsi que l'étude de son évolution dans le temps. Deuxièmement cela permet de rechercher et d'analyser le cycle de vie d'un navire ou de manière agrégée d'un type de navires ou encore d'un pavillon.

Deux métriques sur graphe ont été implémentées ; Elles permettent l'analyse du réseau maritime et de fait favorisent la compréhension de cycles de vie des navires, leurs régularités

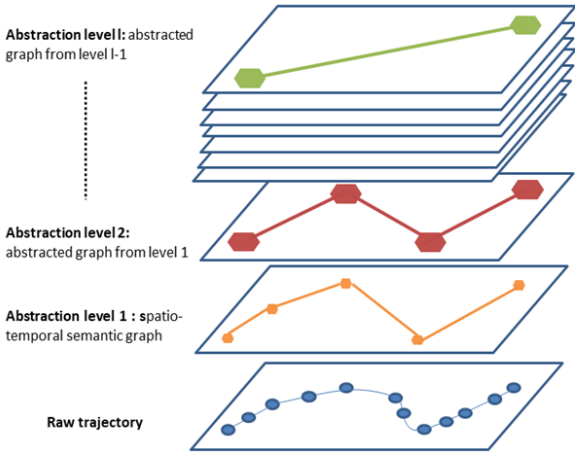


FIG. 2 – Modèle de graphe hiérarchique

ou comportements inhabituels au regard du trafic mondial. Premièrement, l'accessibilité des nœuds. L'accessibilité correspond aux facilités disponibles pour atteindre un nœud en prenant en compte le coût dû au franchissement de la séparation spatiale avec les autres nœuds. Nous avons ainsi proposé, implémenté et évalué une accessibilité hiérarchique compatible avec le modèle de graphe. Deuxièmement, nous nous sommes intéressés à l'évolution dynamique du graphe. Pour cela nous avons défini une mesure de similarité pour évaluer les modifications du graphe dans le temps. La mesure prend en compte la hiérarchie du modèle de graphe et la différence de granularité temporelle entre les intervalles de temps évalués.

3 Gestion des données spatio-temporelles maritimes

Les informations de localisation des navires sont indispensables à une démarche relevant de l'intelligence géospatiale maritime. Dans ce contexte, la gestion des données historiques est aussi importante que l'intégration des flux de données temps-réel. Afin de gérer, et de requêter ces données nous avons développé une chaîne de traitement permettant de gérer le volume et de les mettre à disposition pour les différents cas d'usage. Nous avons opté pour une architecture basée sur la suite ELK composée de Elastic Search (base de données NoSQL), Logstach (utilitaire de collecte de log) et Kibana (Outils de supervision, d'exploration et d'analyse de base de données) qui alimente des développements en Python et une base de données graphe conçue pour l'analyse du réseau maritime et le cycle de vie des navires (cf. Figure 3).

Dans cette chaîne de traitement, les données de localisation historiques (cf. section 2.1) et temps-réel sont intégrées dans la base ELK. Similairement les données dites externes utiles à la construction des cas d'usage (zones géographiques, conditions environnementales, informations documentaires, etc.) peuvent être intégrées à la base de données. Enfin nous y associons une base de données portuaire mondiale (environ 21 000 entités) comprenant la localisation du port, son nom et son extension spatiale sous forme hexagonale issus du clustering de données AIS historiques (Ménard et al., 2021). Du point de vue technique, la suite ELK est instan-

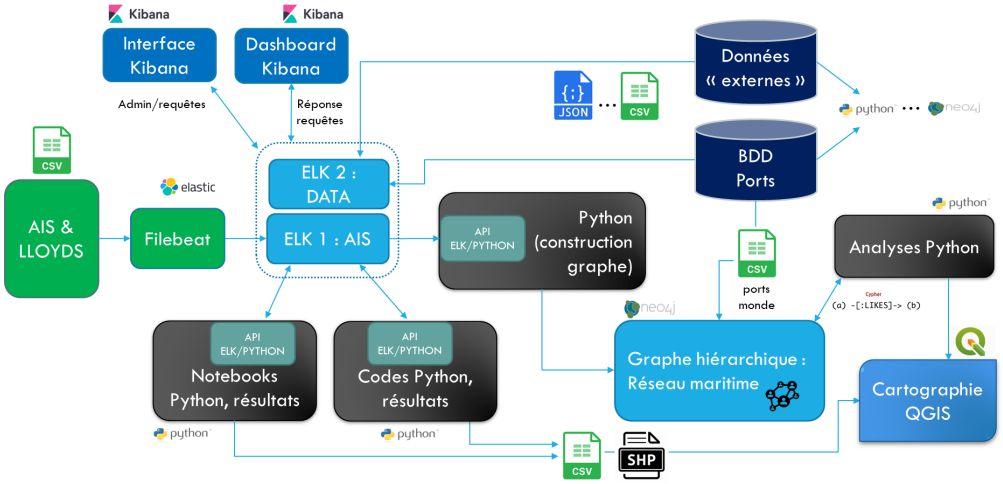


FIG. 3 – Chaîne de traitement pour l'intelligence géospatiale maritime

ciée (sans répliquas) dans 3 conteneurs où chaque conteneur dispose de 64Go de mémoire, 30 processeurs, un volume de données et 30 *shards*.

Le modèle de graphe hiérarchique a été mis en œuvre et testé sur les données réelles de trajectoires maritimes. L'implémentation combine plusieurs outils tels que Python, la bibliothèque MovingPandas [15], l'index H3 d'Uber, le système de gestion de base de données de graphes Neo4j (Community Edition disponible sous licence GPLv3) et son langage d'interrogation Cypher (Maslek Elayam et al., 2022). L'algorithme fonctionne en trois grandes étapes. Après une étape de pré-traitement (eg. suppression d'éventuels doublons à vitesse nulle), les trajectoires sont extraites à l'aide de MovingPandas. Les trajectoires sont ensuite annotés avec les index des hexagones traversés puis compression pour ne retenir que les points remarquables des trajectoires (i.e. point tournants, origines et destinations). Ces points sont ensuite agrégés avec HDBSCAN pour définir les noeuds du graphe. Cette structure peut ensuite être requêtée et les résultats (cycle de vie des navires, sous-graphe, mesure de centralité, etc.) peuvent être utilisé par l'analyste pour enrichir ou annoter sa production cartographique.

Kibana est un plug-in de visualisation de données développé pour permettre la création de visualisations à partir de bases de données ElasticSearch. Kibana est conçu pour être utilisé à partir d'un navigateur web, et permet tant d'explorer des jeux de données de manière approfondie à l'aide d'une interface graphique facile d'utilisation que de créer des visualisations utiles à l'interprétation ou pour compléter les cartes produites à l'aide du logiciel QGIS. Kibana permet notamment de créer des dashboards automatisés, qui peuvent aussi bien inclure les visualisations statistiques habituelles d'un dashboard (histogrammes, série temporelles, graphes, "donuts", etc...) que des cartes intégrant de l'information géospatiale. La figure 4 illustre un exemple simple dans lequel une carte présente certains navires en mer de Chine et trois graphiques présentent les nombres de présence de ces navires et leur répartition durant l'année.

La partie analyse est pour l'essentiel réalisée en Python, notamment au travers de *notebooks* permettant l'extraction, le traitement est l'analyse des données nécessaires aux diffé-

15. https://github.com/anitagraser/movingpandas.

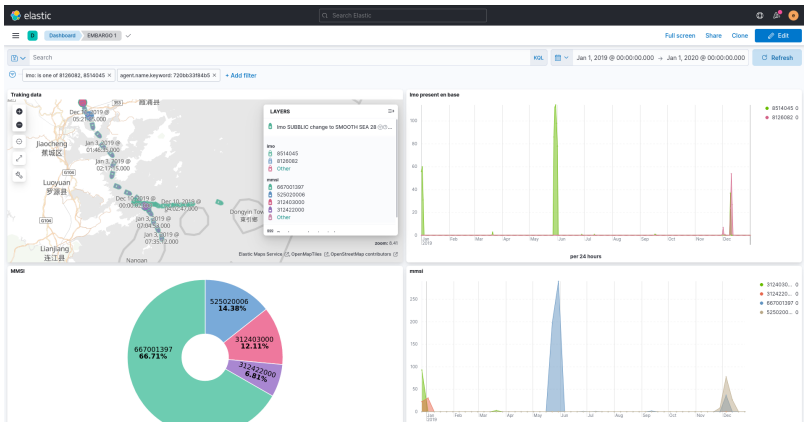


FIG. 4 – Exemple de dashboard Kibana

rentes études. Les développements des algorithmes, méthodes et études réalisés sont guidés par le besoin d'analyse. Actuellement nous avons mis en oeuvre quelques développements communs à plusieurs études sur le prétraitement des données AIS, l'exploitation des hexagones, etc. D'un point vu algorithmique nous avons également réalisé quatre fonctionnalités basées sur les données AIS historiques :

Analyse des flux maritimes : afin de connaître le flux maritime journalier, par exemple, port à port ; mer à mer ou zone économique à zone économique, nous avons conçu un algorithme générique permettant de quantifier les entrées et les sorties dans une zone et applicable sur n'importe quel espace maritime. L'algorithme développé permet de produire des tableaux de bord des flux maritimes entre zones géographiques. Le résultat est généralement présenté sous forme tabulaire ou de *chord diagram* en fonction des besoins (Figure 7).

Le changement d'identité : l'algorithme développé vise à détecter les disparitions et réapparitions de navires accompagnées d'un changement d'identité (susceptible d'être illicite). Pour cela, lorsqu'un navire cesse d'émettre, nous avons développé une expansion hexagonale multi-cibles (multi-threadé d'un point de vue informatique) avec une loi d'évolution gaussienne adaptée au type de navire. La couverture hexagonale s'étend avec le temps. L'apparition de navires dans ces bulles d'expansion déclenche la recherche de falsification d'identité (Figure 8).

L'étude de la densité et de la couverture AIS : un des freins à la parfaite étude de la situation maritime de surface est la présence de zone d'ombres susceptibles de masquer des activités maritimes et qui rend difficile l'identification de la malversation du système (Salmon et al., 2016). Nous avons ainsi conçu et implémenté deux algorithmes. Le premier algorithme agrège les informations historiques de localisation pour produire des cartes de densité (Figure 6). Le résultat est un ensemble d'hexagones (de taille paramétrable), chacun annoté avec des informations statistiques (nombre de navires, vitesse moyenne, etc.). En complément un algorithme d'apprentissage (basé sur XGBoost) pour la prédiction de la couverture AIS d'un récepteur a été conçu. L'algorithme utilise les données AIS historiques couplées aux données

environnementales collectées par le programme Européen Copernicus. Les données sont alignées spatiallement et temporellement grâce aux hexagones H3. Les cartes résultantes (période paramétrable) sont également des cartes d'hexagones.

La prédiction de trajectoires : dans ce travail nous avons souhaité explorer la pertinence d'algorithmes d'apprentissage pour la prédiction de trajectoires représentées à l'aide d'hexagones H3. Nous avons opté pour une approche originale qui transforme les indexes de la séquence d'hexagones d'une trajectoire en caractère de langue naturelle (le chinois en l'occurrence). Nous avons ensuite utilisé le modèle RoBERTa (Robustly Optimized BERT Pretraining Approach) sur ces textes pour faire de la prédiction de trajectoires. La prédiction est une séquence d'hexagones (Figure 5).

FIG. 5 – Exemple de suivi de route courbe par le modèle prédictif. Jaune : hexagones donnés au modèle, rouge : trajectoire réelle, vert : prévision

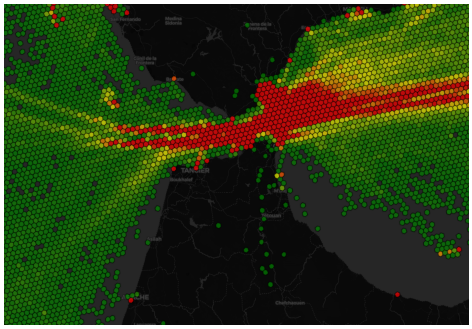


FIG. 6 – Représentation hexagonale de la densité normalisée du trafic maritime au Détroit de Gibraltar sur un mois (janvier 2016)

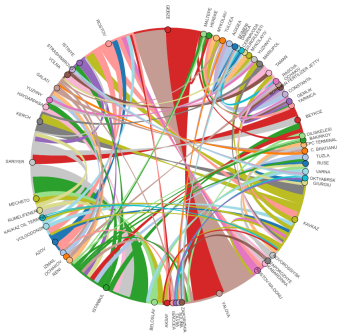


FIG. 7 – Flux maritimes entre les ports de la Mer noire sur une journée (juin 2021)

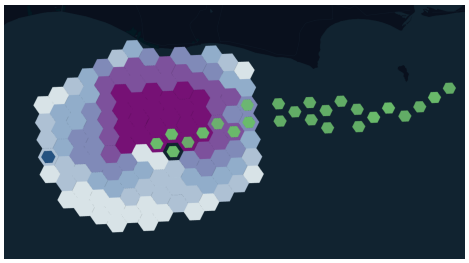


FIG. 8 – Estimation de l'expansion hexagonale au bout de 21 minutes après la disparition d'un cargo (trajectoire verte) en mer de Chine (hexagone de résolution 8 : arrête de 461 mètres)

4 Étude des enjeux des câbles sous-marins en zone Indo-Pacifique

Les rapports de la National Geospatial Intelligence Agency (NGIA) attribuent à l'intelligence géospatiale maritime un nombre restreint d'objets d'étude : cartographies nécessaires à la navigation, planification des opérations maritimes (NGIA, 2006) et sécurité durant la navigation (NGIA, 2018). Pourtant, au regard de la littérature spécialisée, nous avons identifié huit domaines de l'intelligence géospatiale maritime (cf. Figure 9) : les activités de pêches illégales, la sécurité environnementale (lutte contre les pollutions sauvages, préservation des ressources marines), les incidents maritimes, les migrations illégales, le trafic de stupéfiants, la piraterie, le terrorisme et enfin le développement du numérique (interaction entre les activités humaines et les infrastructures de communication). C'est ce dernier axe que nous proposons de développer dans cette section : les enjeux relatifs aux câbles sous-marins.

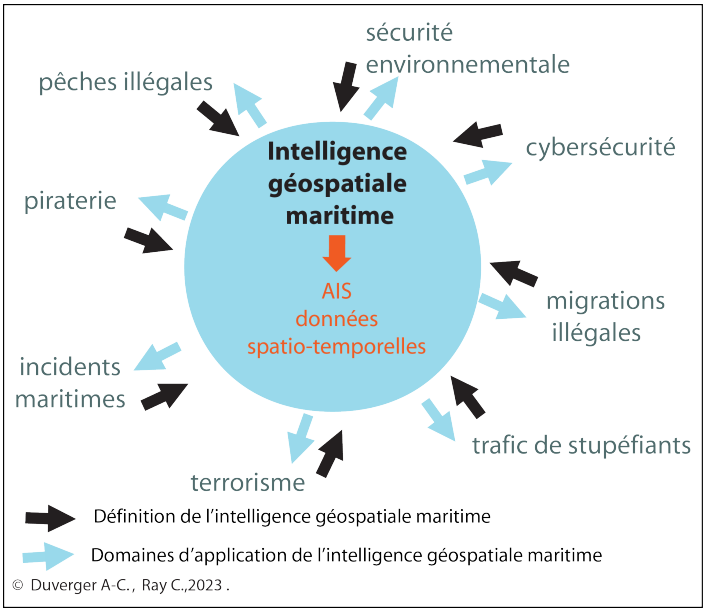


FIG. 9 – Les domaines d'analyse de l'intelligence géospatiale maritime

4.1 Objet et zone d'étude

Les principales artères de communication sont au nombre environ de 450 (Morel, 2023) et permettent à 95% / 97% de nos communications de transiter (Bivaud, 2016). Les vulnérabilités potentielles de ces infrastructures sont de deux ordres : des cyber-vulnérabilités (de type électronique, radio, ou sur l'attaque des protocoles réseaux tel que BGP) et des vulnérabilités physiques (Galland, 2010). En effet, leur endommagement peut conduire à la déstabilisation d'un État ou d'une région puisque l'ensemble de nos activités sont aujourd'hui dépendantes

de réseaux numériques : services administratifs ou bancaires, secteur des transports, du commerce, de la santé, de l'enseignement ou de la recherche (Cattaruzza, 2021).

Nous abordons ce thème sur la zone Indo-Pacifique en raison de deux faits : son caractère hautement maritime et le basculement des enjeux géopolitiques mondiaux vers cette partie du monde. Cependant, ce vaste espace fait l'objet de délimitations variables. Dans le cas de la France, ces dernières adoptent une vision périmétrique à partir de ses sols (Lechevry, 2019). Ainsi, en considérant les espaces d'outre-mer comme les sièges de la projection française en Indo-Pacifique, nous pouvons diviser cet espace en quatre zones : l'Indo-Pacifique occidental, central, austral et oriental. L'étude actuellement menée se concentre sur sa partie orientale, car l'état de la recherche en géographie a montré que les transports maritimes ont peu été analysés dans cette région [16].

4.2 Traitements algorithmiques, indexations hexagonales des données spatio-temporelles et données contextuelles

Afin d'analyser la vulnérabilité des câbles sous-marins pour la partie orientale de l'Indo-Pacifique, plusieurs traitements algorithmiques sont nécessaires (cf. figure 6). Le premier s'appuie sur la dimension temporelle des données AIS ; il s'agit de les filtrer selon une période précise. Pour le présent cas d'étude, des données AIS sont sélectionnées afin d'effectuer des corrélations avec le dernier état des lieux du réseau de câbles (datant de 2018). Notons que cette analyse est centrée sur les territoires français de la zone.

1. Premièrement un filtrage temporel est réalisé ; le mois de juillet 2019 dans notre cas (un mois hors de la période cyclonique ; cette dernière pouvant impacter les activités des navires).
2. La seconde étape vise à sélectionner les données d'une zone géographique précise, qui peut être définie selon deux variables : les coordonnées géographiques et par le biais des délimitations des mers et océans du monde (données issues du *Flanders Marine Institute*).

Éventuellement, en fonction des cas d'usage, une étape de prétraitement des données est possible (par exemple ; ne conserver que les navires en mouvement). Ensuite, les données AIS sont intégrées dans des hexagones Uber H3. Ces derniers concernent les zones à proximité des câbles, permettant de créer « *un profil de risque le long du câble* » (Doan et al., 2016). Pour cette étude à l'échelle régionale, des hexagones d'une résolution de 4 ont été choisis (arête d'une longueur moyenne de 22,6 km) (Maslek Elayam et al., 2022). Ainsi, au regard de l'échelle de la carte : 1/400 km, nous obtenons un niveau de détail suffisant, tout en agrégeant un volume important d'informations.

3. La troisième étape consiste à agréger dans les hexagones les navires présents, selon leur type (cargos, navire militaires, remorqueurs, voiliers…).
4. La quatrième étape agrège les positions des navires câbliers. Elle est basée sur la liste de ces navires établie par l'*International Cable Protection Commitee*.

16. Soit les espaces maritimes compris entre le 40^{ime} parallèle nord et le 60^{ime} parallèle sud ; entre le 100^{ime} parallèle est et le 60^{ime} parallèle ouest.

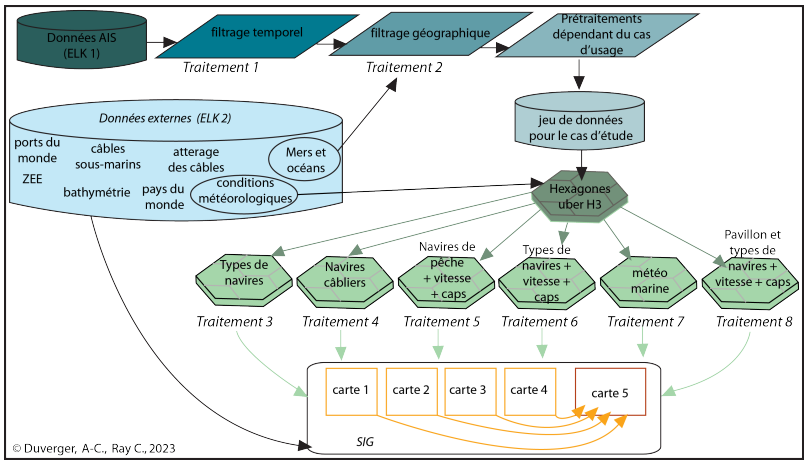


FIG. 10 – L'intelligence géospatiale appliquée à l'analyse des câbles sous-marins en Indo-Pacique

5. La cinquième étape permet de construire des hexagones contenant seulement des informations relatives aux navires de pêche (nombre, vitesse et caps : cap vrai et cap suivi sur le fond).
6. La sixième étape vise à obtenir des renseignements sur les divers types de navires, mais également sur leur vitesse et leurs caps.
7. Ensuite, un autre traitement génère dans les hexagones des données météorologiques : vagues (hauteur) et vents (vitesse) selon leur direction (données produites par le programme *Copernicus* de l'Union européenne).
8. Enfin, le huitième traitement de ces données spatio-temporelles permet de construire des hexagones contenant à la fois les types de navires présents, leurs pavillons ainsi que leur vitesse et leurs caps.

Afin d'enrichir l'analyse, des données contextuelles sont nécessaires : celles des câbles sous-marins et leurs terminaux (produites par *Telegeography* [17]) concernent l'année 2018. Les mers et océans du monde ; des données constituées par la *Flanders Marine Institute* (2018). Les informations relatives aux ZEE ont aussi été produites par la *Flanders Marine Institute* et la *Vlaams Instituut voor de Zee* (2018). Les frontières administratives des pays (*World Food Programme*, 2022) sont aussi indispensables à l'analyse. Les données relatives aux ports et zones stationnaires mobilisées sont produites par des chercheurs ayant fusionné quatre bases de données : celle du *World Port Index* (*World Food Programme*) ; les mouillages de *Global Fishing Watch* et la base de ports de l'IFREMER [18] (Menard et al., 2021). Concernant la bathymétrie, deux jeux de données sont utilisés : celui de la *North American Cartographic Information Society* (2009) et celui de l'OHI et de la COI [19].

17. Une société spécialisée dans le marché des télécommunications
18. Institut français de recherche pour l'exploitation de la mer
19. Organisation hydrographique internationale et Commission océanographique intergouvernementale

4.3 Projets cartographiques et étude de cas

Les traitements algorithmiques générant l'indexation hexagonale des données AIS (sous la forme de fichier Shapefile) ainsi que l'ajout de données ouvertes dans le SIG, permettront la réalisation de cartes relatives à la vulnérabilité des câbles sous-marins :

— densité des navires et vulnérabilité des câbles sous-marins.
— activités de pêche et vulnérabilité des câbles sous-marins.
— ancrage des navires et vulnérabilité des câbles sous-marins.
— actes de malveillance et vulnérabilité des câbles sous-marins.

Puis, un indice de vulnérabilité des câbles sera créé. Celui-ci sera représenté sous la forme de diagrammes radiaux. Ces derniers comporteront six dimensions et seront associés à chaque hexagone. Les différentes dimensions des diagrammes sont construites en associant des valeurs statistiques connues (par exemple, 44 % des incidents sur les câbles sont liés aux activités de pêche et 14 % à des problèmes d'ancrage des navires) (ICPC, UNEP, WCMC, 2009) à des informations relatives aux navires présents et à leur environnement (densité de navires, bathymétrie, météorologie marine, nombre de câbles par pays et temps d'intervention des navires câbliers). Ces dernières varient selon la zone d'étude considérée, rendant cette méthode d'identification de zones à risques pour les câbles sous-marins, applicable à d'autres zones géographiques.

Enfin, nous nous concentrerons sur une rupture de câble avérée : le 10 mai 2023 aux Îles Salomon. Une étude de cas sélectionnée, car elle correspond à une situation dominante : la majorité des endommagements de câbles sont liés à des activités de pêche (ICPC, UNEP, WCMC, 2009). De plus, les Iles Salomon sont un territoire insulaire [20]. En ce sens, elles entretiennent des caractéristiques communes avec les collectivités françaises [21] de la zone qui, du fait de leur multi-insularité, sont généralement plus vulnérables aux pannes d'Internet liées aux câbles sous-marins, car elles n'ont pas accès à des réseaux câblés terrestres denses (Bueger et al., 2022).

Cette situation d'endommagement de câble permettra de prévenir des événements similaires. Pour cela, nous nous baserons sur le comportement du navire impliqué, sur ses caractéristiques de circulation et l'environnement maritime dans lequel il évoluait au moment de l'incident. Nous travaillons à partir de données AIS relevant de trois filtrages temporels différentes :

— les 12 mois précédents l'incident nous renseigneront sur le comportement routinier du navire (mouillages privilégiés, trajets réguliers, zones de pêche...) ;
— l'heure qui a précédé et celle qui a suivi la rupture du câble permettront de décrire avec précision les conditions de survenue de celle-ci ;
— enfin, le mois suivant l'incident est considéré afin de rendre compte des événements ayant engendré l'arrivée tardive du câblier [22].

Pour apprécier l'environnement maritime du navire concerné, nous nous appuyons sur la localisation des câbles sous-marins ; ce jeu de données comprend des informations indispensables à une approche géopolitique (date d'installation et propriétaire du câble). Les données

20. Environ 1000 îles (MACBIO, 2019) dont 164 sont habitées (SINSO, 2009).
21. La Nouvelle-Calédonie compte environ 141 îles dont 15 sont peuplées (Duverger, 2021) ; la Polynésie française comprend 121 îles dont 72 sont habitées (ISPF, 2015).
22. Ce dernier n'est arrivé sur place que le 20 juin (DCmag, 2023).

relatives aux ZEE et aux frontières des pays nourrissent aussi ce type d'analyse, car l'extrémité d'un câble se situe généralement dans d'autres pays (O'Malley, 2019).

Les zones d'atterrages sont également intégrées, car si elles sont proximité les unes des autres, il existe un risque accru de défaillance dû à une même perturbation (Bueger et al., 2022). Ainsi, on pourra caractériser la gravité de l'événement. Les données relatives aux ports et zones stationnaires permettent d'analyser les interactions entre les situations statiques et l'événement. Nous exploitons également des la bathymétrie pour enrichir notre examen de la situation du câble. Les données de vitesse du vent et de hauteur des vagues (selon leur direction) le 10 mai 2023 sont intégrées, afin d'approcher l'influence des conditions météorologiques sur la trajectoire du navire. Enfin, cette étude demande de recourir à des hexagones de résolution 11 (arrête d'une longueur moyenne de 24.9 m) (Maslek Elayam et al., 2022), pour rendre compte de la densité de navires à une grande échelle, et ce, pour le filtrage temporel proche de l'heure de l'incident.

Conclusion

Nous avons mis en évidence qu'en raison du basculement des enjeux géopolitiques mondiaux vers la zone Indo-Pacifique, l'intelligence géospatiale doit aujourd'hui se tourner vers l'analyse de l'espace maritime. Pour cela elle dispose de diverses sources d'informations associées aux quatre dimensions physiques du milieu maritime, notamment les données AIS. Cependant, en raison de leur caractère massif (géoréférencées, horodétées et enrichies d'informations statistiques), ces dernières appellent à penser des modes de gestions, de représentation et d'analyse spécifiques, nécessitant des compétences qui relèvent aussi bien des sciences de la donnée, de la géomatique que de la géographie.

Ainsi, du point de vue de la gestion et de l'analyse de ces données spatio-temporelles maritimes, nous avons montré l'intérêt de la construction d'une infrastructure basée sur la suite Elastic Search associée à des développements en Python. Ils permettent de procéder à la création de graphes et d'algorithmes originaux indispensables à la représentation et à la compréhension des mobilités et flux maritimes. Par ailleurs, l'étude de la vulnérabilité des câbles sous-marins confirme le potentiel de l'indexation hexagonale pour spatialiser des informations diverses (densité, vitesses, pavillons des navires...).

Enfin, nous avons pu constater que du fait des détournements inhérents à la nature de ces données, les algorithmes développés permettent d'affiner les analyses en détectant les anomalies et en offrant la possibilité de reconstruire les trajectoires des navires. Cela ouvre de nombreuses perspectives d'études dans divers domaines de l'intelligence géospatiale : du terrorisme à l'examen des pêches illégales.

Références

Abu Sari, N., M. Y. B. Abu Sari, A. Ahmad, S. Sahib, et F. Othman (2018). Using laper quadcopter imagery for precision oil palm geospatial intelligence (op geoint). *Electronic and Computer Engineering 10*(1), 25–33.

Bivaud, A. (2016). Les dommages causés aux câbles sous-marins.

Boulanger, P. (2006). *Geographie militaire*. Carrefours, Ellipses.

Boulanger, P. (2011). Renseignement géographique et culture militaire. *Hérodote n°140*, 47–63.

Boulanger, P. (2019a). Le geoint à l'origine d'une nouvelle science de l'information géospatiale ? *Défense n°200*(2-3), pp.32–36.

Boulanger, P. (2019b). Le geospatial intelligence français : quels enjeux et défis aujourd'hui ? *Défense et stratégie hors-série*(63), 70–76.

Boulanger, P. (2020). *La géographie reine des batailles*. Perrin, Paris.

Bueger, C., T. Liebetrau, et J. Franken (2022). Security threats to undersea communications cables and infrastructure – consequences for the eu. Technical report, European Parliament and Bueger Christian and Liebetrau Tobias and Franken Jonas.

Cattaruzza, A. (2021). Vers une géopolitique numérique. *Constructif 3*(60), 46–50.

CICDE (juillet, 2021). Renseignement géospatial, doctrine interarmées-dia-2.17. Technical report, Centre interarmées de concepts, de doctrines et d'expérimentations, Ministère des Armées.

DCmag (Mis en ligne en juin 2023, modifié en août 2023). Le navire a arraché un câble sous-marin en mer de corail, capitaine et propriétaire feront face à la justice. `https://dcmag.fr/le-navire-a-arrache-un-cable-sous-marin-en-mer-de-corail`.

Doan, H., L. Macnay, A. Savadogo, et K. Smith (2016). Profondeur d'enfouissement des câbles offshore par une approche basée sur le risque. *Journées Nationales de Géotechnique et de Géologie de l'Ingénieur*, 1–8.

Ducruet, C. (2013). Network diversity and maritime flows. *Journal of Transport Geography 30*, 77–88.

Ducruet, C. (2015). *Maritime Networks : Spatial Structures and Time Dynamics*, Chapter Co-evolutionary dynamics of ports and cities in the global maritime network, 1950–90, pp. 351–373. Routledge Studies in Transport Analysis.

Ducruet, C. (2016). La spatialité des réseaux maritimes : Contributions maritimes à l'analyse des réseaux en géographie.

Duverger, A.-C. (2021). Systèmes de transports et de mobilités inter-îles en nouvelle-calédonie et au vanuatu.

Fremont, A. et A. Fremont-Vanacore (2014). *Géographie des espaces maritimes*. La documentation française, Dossier n°8104, Paris.

Galland, J.-P. (2010). Critique de la notion d'infrastructure critique. *Flux 3*(81), 6–18.

Gehlen, M., M. R. Nicola, E. R. Costa, V. K. Cabral, E. L. de Quadros, C. O. Chaves, R. A.Lahm, A. D. RNicolella, M. L. Rossetti, et D. R. Silva (2019). Geospatial intelligence and health analitycs : Its application and utility in a city with high tuberculosis incidence in brazil. *journal of Infection and Public Health 12*, 681–689.

ICPC-UNEP-WCMC (2009). Submarine cables and the oceans : connecting the world. Technical report, United Nations Environment Program and World Conservation Monitoring Centre and International Cable Protection Committee.

Iphar, C., A. Napoli, et C. Ray (2015). Detection of false ais messages for the improvement of maritime situational awareness. *OCEANS 2015 MTS/IEEE Washington*, 1–7.

ISPF (2015). Polynésie française en bref. Technical report, Institut de la Statistique de la Polynésie française.

Lechevry, C. (2019). La place des outre-mer océaniens dans la politique indo-pacifique de la france. *Revue de Défense Nationale 8*(823), 21–27.

MACBIO (2019). Marine atlas maximizing benefits for solomon islands. Technical report, Marine and Costal Biodiversity Management in Pacific Island Countries, Deutsche Gesellschaft für Internationale Zusammenarbeit and GRID-Arendal and International Union for Conservation of Nature and Republic of Solomon Islands and Secretariat of the Pacific Regional Environment Programme.

Markogiannaki, O., A. Karavias, D. Bafi, D. Angelou, et I. Parcharidis (2020). A geospatial intelligence application to support post-disaster inspections based on local exposure information and on co-seismic dinsar results : the case of the durres (albania) earthquake on november 26, 2019. *Natural Hazard 103*, 3085–3100.

Maslek Elayam, M., G. Kerhoas, V. Lambert, C. Ray, et A. Ménard (2022). On the interest of hexagonal abstraction of maritime information. In *OCEANS 2022, Hampton Roads, Hampton Roads, VA, USA*.

Maslek Elayam, M., C. Ray, et C. Claramunt (2021). Modèle de graphe pour l'analyse des structures de trajectoires maritimes. In *21ème édition de la conférence Extraction et Gestion des Connaissances (EGC)*.

Maslek Elayam, M., C. Ray, et C. Claramunt (2022). A hierarchical graph-based model for mobility data representation and analysis. *Data and Knowledge Engineering 141*, 102054.

Menard, A., V. L. de Cursay, C. Ray, M. E. M. Charles Guenois, et M. Dréau (2021). Construction d'une méta-base de ports à l'échelle mondiale. *SAGEO-2021, poster*, 1–70.

Miossec, A. (2014). *Géographie des mers et océans*. Presses universitaires de Rennes.

Morel, C. (2023). *Les câbles sous-marins*. Biblis, CNRS Edition, Paris.

Morel, E. et P. Boulanger (2016). *Géographie et guerre. De la géographie militaire au Geospatial Intelligence en France (XVIIIe-XXIe siècle)*, Chapter La géolocalisation et le GEOINT comme outil d'analyse sociétales et géopolitique, pp. 189–197. Bulletin de la Société de géographie.

Ménard, A., V. Lambert de Cursay, C. Ray, C. Guenois, M. Maslek Elayam, et M. Dréau (2021). Construction d'une méta-base de ports à l'échelle mondiale. In *Conférence internationale de Géomatique et d'Analyse Spatiale*. SAGEO'21.

NGIA (avril 2018). Geospatial intelligence (geoint) basic doctrine. Technical report, National Geospatial Intelligence Agency, National System for Geospatial Intelligence.

NGIA (septembre 2006). Geospatial intelligence (geoint) basic doctrine. Technical report, National Geospatial Intelligence Agency, Office of Geospatial-Intelligence Managment.

O'Malley, S. (2019). Vulnerability of south korea's undersea cable communications infrastructure. a geopolitical perspective. *Korea Observer 50*(3), 309–330.

Parambil, N. V., S. Krishnan, et M. Firoz (2021). *Geo-intelligence for Sustainable Development*, Chapter Geo-intelligence-Based Approach for Sustainable Development of Peri-Urban Areas : A Case Study of Kozhikode City, Kerala (India), pp. 35–52. Springer.

Prazuck, C. (2022). *Le Geoint et les océans*. dans : Boulanger Philippe (Dir.), Géographie, Geoint et opérations, Bulletin de liaison des membres de la Société de géographie, Hors-série, pp. 82-89.

Royer, P. (2012). *Géopolitique des Mers et des Océans*. PUF, Paris.

Salmon, L., C. Ray, et C. Claramunt (2016). Continuous detection of black holes for moving objects at sea. pp. 1–10.

SINSO (2009). Basic tables and census description. Technical report, Solomon Islands National Statistical Office.

Tabarly, S. (Mis en ligne en 2005, modifié en mars 2023). Isotropie et anisotropie. `http://geoconfluences.ens-lyon.fr/glossaire/isotropie-anisotropie-1`.

Remerciement

Ce travail est réalisé dans le cadre du projet national GEOINT financé par l'Agence d'Innovation de la Défense (AID).

Summary

Monitoring, analyzing and understanding maritime mobility is a major concern of modern society, as it reflects the commercial, societal and geopolitical stakes that exist between countries. The volumes of data describing these maritime mobilities are growing and fluctuating, and the extraction and management of knowledge for the benefit of, for example, fisheries monitoring, trade, illegal acts, etc. requires a complementary approach combining geographic information engineering, artificial intelligence and human sciences. In this context, our aim is to develop multi-disciplinary capabilities for analyzing and interpreting maritime mobilities (what are the ships, what are their characteristics, etc.), analyzing their maritime behaviors (where do they move, at what speed, etc.) and analyzing their maritime situations (in what context do they move, why, what is the international context, etc.). In this article, we propose an analysis methodology and present the data management architecture constructed. The work is illustrated by a preliminary study of the maritime digital challenges of the Indo-Pacific region.

Multi-SPMiner : un framework d'apprentissage profond pour l'extraction de motifs fréquents dans des graphes spatio-temporels

Assaad Zeghina*, Aurélie Leborgne*
Florence Le Ber*, Antoine Vacavant**

*Université de Strasbourg, CNRS, ENGEES, ICube UMR 7357, F67000 Strasbourg
{assaad-oussama.zeghina, aurelie.leborgne, florence.le-ber}@icube.unistra.fr,
**Universite Clermont Auvergne, CNRS, SIGMA Clermont, Institut Pascal
antoine.vacavant@uca.fr

Résumé. Cet article présente un nouveau framework, Multi-SPMiner, pour l'extraction de motifs fréquents dans les multigraphes, en particulier les graphes spatio-temporels. Il utilise une approche en deux étapes pour extraire ces motifs, 1) en plongeant les nœuds dans un espace ordonné et 2) en faisant croître itérativement les motifs en parcourant cet espace. Les expérimentations menées sur des jeux de donnes synthétiques démontrent l'efficacité de Multi-SPMiner pour identifier des motifs fréquents dans des multigraphes. De plus, il est testé sur des graphes simples, montrant sa généralité par rapport à SPMiner. Ce texte est un résumé d'un article présenté à la 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES2023).

1 Introduction

Les graphes sont utilisés pour représenter des données relationnelles de différents domaines, notamment les réseaux sociaux, la biologie et les systèmes de transport (Zhang et al., 2020). A la disponibilité croissante de telles données, correspond un besoin croissant pour des algorithmes efficaces permettant d'extraire des motifs portant des informations précieuses. L'extraction des motifs fréquents, également connue sous le nom d'extraction de sous-graphes, est une tâche majeure dans l'analyse de graphes, permettant d'identifier des sous-graphes récurrents dans un grand graphe ou plusieurs graphes (Jiang et al., 2013). Ces sous-graphes révèlent des motifs structurels offrant des informations sur les relations et les propriétés des données, avec par exemple des applications en biologie, en sciences sociales et en chimie (Fournier-Viger et al., 2020). Des algorithmes tels que SIGRAM (Kuramochi et Karypis, 2005), GRAMI (Elseidy et al., 2014) et MuGram (Ingalalli et al., 2018) ont été conçus pour l'extraction fréquent de motifs dans les graphes simples.

Pour la modélisation des systèmes et des relations complexes, où les entités entretiennent des interactions diverses, par exemple dans les réseaux sociaux, les multigraphes offrent une représentation plus complète. Mais peu d'algorithmes de recherche de motifs fréquents sont

adaptés à ces données. MuGram (Ingalalli et al., 2018) permet de traiter des multigraphes, mais reste coûteux en calcul et en mémoire. Par conséquent, il existe un besoin d'algorithmes adaptés pour l'extraction efficace de motifs fréquent dans les multigraphes. Les approches existantes pour des structures de graphes plus complexes ont des limitations pour capturer pleinement les complexités des multigraphes. Par ailleurs, la croissance combinatoire des ensembles de candidats lors de l'extraction de motifs dans les graphes, en particulier pour des grands motifs, justifie de considérer des approches d'apprentissage profond. Une telle approche est exploitée dans SPMiner (Ying et al., 2020) qui utilise un réseau neuronal de graphes et recherche les motifs via un parcours de l'espace de plongement. SPMiner est originellement conçu pour traiter des graphes simples. Dans ce papier, nous étendons SPMiner aux multigraphes étiquetés et dirigés, en mettant l'accent sur les graphes spatio-temporels. Notre objectif est d"extraction de motifs de façon efficace dans ces structures de graphes complexes. Nous évaluons notre technique sur des graphes spatio-temporels synthétiques, démontrant son potentiel pour diverses applications.

2 Définitions

Definition 1 (Multigraphe)**.** *Un multigraphe $G = (V, E, l_V, l_E, L_V, L_E)$ est une structure de données composée d'un ensemble V de sommets ou de nœuds, d'un ensemble d'arêtes $E \subseteq V \times V$, de deux ensembles d'étiquettes L_V, L_E, et de deux fonctions, $l_V : V \to 2^{L_V}$, qui associe un sous-ensemble d'étiquettes à chaque sommet, et $l_E : E \to 2^{L_E}$, qui associe un sous-ensemble d'étiquettes à chaque arête.*

Definition 2 (Sous-graphe)**.** *Un sous-graphe d'un multigraphe $G = (V, E, l_V, l_E, L_V, L_E)$ est un multigraphe $G' = (V', E', l_{V'}, l_{E'}, L_V, L_E)$ tel que $V' \subseteq V$, $E' \subseteq E$, et $\forall v \in V', l_{V'}(v) \subseteq l_V(v)$ et $\forall e \in E', l_{E'}(e) \subseteq l_E(e)$. G' est un sous-graphe induit par les nœuds si et seulement si : $V' \subseteq V$, $E' = \{(u, v) \in E \mid u \in V', v \in V'\}$. G' est un sous-graphe induit par les arêtes si et seulement si $E' \subseteq E$, et $V' = \{v \in V \mid \exists u \in V \text{ et } (u, v) \in E'\}$.*

Definition 3 (Isomorphisme de sous-graphe)**.** *Étant donné $G = (V_G, E_G, l_{V_G}, l_{E_G}, L_V, L_E)$ et $H = (V_H, E_H, l_{V_H}, l_{E_H}, L_V, L_E)$ deux multigraphes, nous disons que H est isomorphe à un sous-graphe de G, noté par $H \subseteq G$, s'il existe une application bijective $\phi : V_H \to V_G$ telle que $(u, v) \in E_H$ si et seulement si $(\phi(u), \phi(v)) \in E_G$, $l_{V_H}(u) \subseteq l_{V_G}(\phi(u))$, $l_{V_H}(v) \subseteq l_{V_G}(\phi(v))$ et $l_{E_H}(u, v) \subseteq l_{E_G}(\phi(u), \phi(v))$.*

Graphes spatio-temporels. Les graphes spatio-temporels (st-graphes), tels qu'introduits par (Del Mondo et al., 2013), sont constitués d'entités temporelles et de leurs relations, qui sont des relations spatiales, spatio-temporelles et de filiation. Dans ce modèle une relation relie toujours deux entités soit au même instant de temps, soit à deux instants de temps consécutifs. Ainsi, un st-graphe est défini sur un domaine temporel, $\mathcal{T} = \{t_1, t_2, \dots t_n\}$, où t_i représente un instant de temps d'une granularité donnée et $t_i < t_{i+1}$ pour tout $i \in [1, n]$. $\Delta = \{e_1, e_2, \dots e_m\}$ est un ensemble d'entités. Nous introduisons également Σ, un ensemble de relations spatiales, et Φ, un ensemble de relations de filiation (Leborgne et al., 2021a).

Definition 4. *Un graphe spatio-temporel est un multigraphe $G = (V, E, l_V, l_E, L_V, L_E)$, où l'ensemble des étiquettes de sommets $L_V = \Delta \times \mathcal{T}$ et l'ensemble des étiquettes d'arêtes $L_E = \Sigma \cup \Phi$.*

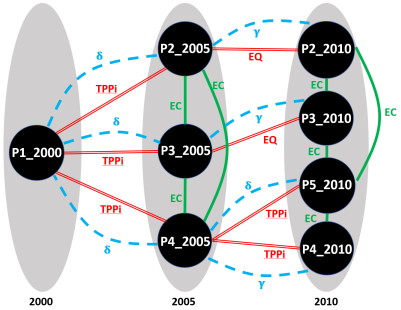


FIG. 1 – *Une représentation st-graphe : lignes rouges pour les relations spatio-temporelles, vertes pour spatiales, et bleues pour la filiation. Leborgne et al. (2021a).*

Un st-graphe peut être vu comme l'union de trois sous-graphes (comme indiqué dans la figure 1) dans lesquels les entités sont regroupées selon les instants de temps. Le sous-graphe spatial montre des interactions spatiales (en vert sur la figure), le sous-graphe spatio-temporel représente les interactions entre les entités à des moments successifs (en rouge), tous deux modélisées par les relations de la théorie RCC8 (Cohn et al., 1997). Le sous-graphe de filiation montre la transmission d'identité, avec une relation de continuation γ signifiant une identité partagée et une relation de dérivation δ dénotant qu'une partie de l'identité d'une entité est contenue dans une autre (en pointillé bleu sur la figure).

Convolution de graphes et convolutions de multigraphes. Les réseaux de convolution de graphes (GCNs) sont des architectures de réseaux neuronaux spécifiquement conçues pour fonctionner sur des données structurées en graphe, permettant l'apprentissage efficace et la représentation des caractéristiques au niveau des nœuds en agrégeant les informations des nœuds voisins. Pour un graphe donné G, soit $\mathbf{X} \in \mathbb{R}^{n \times d}$ la matrice de caractéristiques des nœuds, où d est le nombre de caractéristiques pour n nœuds. Soit $\mathbf{A} \in \mathbb{R}^{n \times n}$ la matrice d'adjacence du graphe G. La couche de convolution du premier ordre d'un GCN calcule la représentation cachée $\mathbf{H} \in \mathbb{R}^n \times d'$ (d' est la taille des caractéristiques résultant du plongement des nœuds) comme suit :

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \tag{1}$$

où $\mathbf{H}^{(l)}$ représente la l-ème couche du GCN (*et* $H^(0) = X$ *?*, $\tilde{\mathbf{D}}$ est la matrice de degré de $\tilde{\mathbf{A}}$, et $\mathbf{W}^{(l)}$ est la matrice de poids de la l-ème couche. La fonction σ est typiquement une fonction d'activation non linéaire telle que ReLU.

Les réseaux de convolution de multigraphes (MGCN) sont une généralisation des GCNs pour gérer plusieurs graphes. La couche de convolution de multigraphes fonctionne sur chaque graphe individuellement, et les résultats sont combinés pour produire une sortie finale.

3 Méthode proposée

Dans cette section, nous présentons Multi-SPMiner, une extension de SPMiner (Ying et al., 2020) adaptée à l'extraction fréquente de motifs dans des multigraphes. Contrairement à la mé-

thode de base axée sur l'identification de motifs de sous-graphes fréquents dans des ensembles de graphes, Multi-SPMiner vise à trouver dans un seul multigraphe tous les motifs ancrés à de nœuds avec une valeur de support supérieure à un seuil fixé s. Il décompose le graphe en voisinages et utilise un MGCN pour obtenir des plongements ordonnés. Ensuite, à partir d'un nœud initial, le framework ajoute de manière itérative des nœuds et des arêtes à un graphe réduit à un nœud pour trouver des motifs fréquents, comme illustré dans la figure 2.

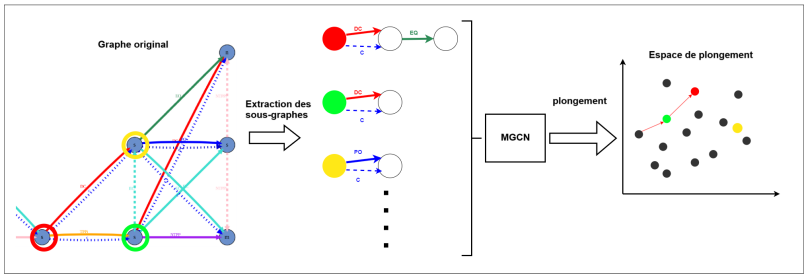


FIG. 2 – *Les deux étapes du framework Multi-SPMiner, inspirées de SPMiner*

Phase de plongement. Dans cette phase, le graphe est d'abord décomposé en voisinages N de k sauts ancrés à chaque nœud v, où K est fixé à une valeur supérieure à la taille maximale du motif pouvant être trouvé dans le graphe (pour un motif de taille 10, k est fixé à 12), comme illustré dans la figure 2.

Les voisinages sont encodés à l'aide d'un réseau de convolution multigraphe dans un espace de plongement ordonné, suivant le principe présenté par Vendrov et al. (2016) et Ying et al. (2020). Le plongement ordonné est une technique qui préserve l'arrangement relatif des éléments, ce qui le rend bien adapté pour modéliser les relations de sous-graphes. Deux graphes A et B sont tels que $A \leq B$ si et seuement A est isomorphe à un sous-graphe de B. Pour imposer la contrainte de plongement ordonné, nous introduisons une fonction de plongement $\phi : \mathcal{G} \mapsto \mathbb{R}^n$ qui projette les graphes sous forme de vecteurs, tels que $A \leq B$ si et seulement si $\phi(A) \leq \phi(B)$. Dans le plan 2D, $\phi(A)$ est alors "plus bas à gauche" de $\phi(B)$, comme illustré dans la Figure 2 par les voisinages ancrés aux nœuds verts et rouges. Le processus d'apprentissage produit un réseau de convolution multigraphe qui préserve l'ordre dans l'espace de plongement.

L'entraînement utilise un ensemble de paires positives (A, B) où A est un sous-graphe de B et des paires négatives (A', B') où A' n'est pas un sous-graphe de B' pour optimiser la perte de marge maximale, définie comme :

$$\sum_{(A,B)\in P} E(A,B) + \sum_{(A',B')\in N} \max\left(0, \alpha - E\left(A',B'\right)\right) \tag{2}$$

où α est un hyperparamètre de marge et E est la pénalité de plongement ordonné définie comme suit :

$$E(A,B) = \| \max(0, \phi(A) - \phi(B))\|^2. \tag{3}$$

Les exemples positifs sont générés en échantillonnant un sous-graphe plus petit à partir d'un voisinage cible aléatoire, tandis que les exemples négatifs sont générés en échantillonnant aléatoirement un sous-graphe différent. Notre GNN se compose de plusieurs couches d'opérations de convolution multigraphe avec des couches de saut pour extraire les attributs structurels de voisinages de tailles variées.

Phase de recherche. L'approche se concentre sur la recherche de motifs ancrés à des nœuds, fréquemment rencontrés dans un graphe donné. Elle utilise les plongements de l'étape MGCN précédente, qui représentent les sous-graphes ancrés à des nœuds dans un espace ordonné. En raison du nombre exponentiel de possibilités, la recherche directe de motifs fréquents est difficile. Pour pallier cela, une procédure de recherche itérative a été choisie : le graphe initial de taille 1 (réduit à un nœud), est sélectionné aléatoirement ; à chaque étape, ce graphe grandit par ajout des nœuds adjacents suivant une marche monotone dans l'espace de plongement ; finalement les motifs sont sélectionnés en fonction de leur fréquence. L'ajout itératif de nœuds vise à minimiser la marge totale $m(G)$, définie comme suit :

$$m(G) = \sum_{N \in \mathcal{N}} |\max(0, \phi(G) - \phi(N))|^2 \tag{4}$$

où $\mathcal{N}$ représente les graphes de voisinages de tous les nœuds dans G, et ϕ est la fonction de plongement ordonné obtenue à partir du MGCN. Contrairement à SPMiner, Multi-SPMiner considère tous les nœuds adjacents lors du processus d'ajout(pour les sous-graphe), opérant sur un sous-graphe candidat à la fois.

4 Expérimentations

Pour entraîner le réseau de convolution multigraphe et évaluer les performances du framework, nous avons utilisé un générateur de graphe spatio-temporel synthétique (Leborgne et al., 2021a) conçu pour imiter des modèles de données réels. Ce générateur produit des st-graphes incluant des motifs fréquents. Nous avons mené 22 expérimentations, chacune utilisant le générateur de données, selon trois configurations, décrites ci-dessous, qui permettent d'évaluer Multi-SPMiner en termes de justesse, de capacité de généralisation et de robustesse. Dans l'analyse des résultats, nous mettons l'accent sur l'exactitude des motifs extraits à partir d'un unique graphe généré. Une comparaison avec SPMiner a également été menée sur des graphes simples.

Configuration 1. Le tableau 1 détaille 14 expériences, chacune impliquant la génération de 200 graphes d'entraînement et 10 graphes de test. Des motifs de tailles différentes ont été insérés dans chaque graphe. Par exemple, le premier test est effectué sur des graphes spatio-temporels avec 200 nœuds, 40 nœuds par temporalité, des distributions de relations fixées par les valeurs [5,5,2] et des motifs de taille 5 ou 6. Dans les tests suivants, on fait varier les paramètres tels que la taille du graphe, la taille du motif et le nombre de motifs, afin d'analyser leur impact sur le taux des motifs obtenus. Les résultats présentés aux tableau 1 indiquent que Multi-SPMiner a atteint des performances stables, avec un taux de reconnaissance (*accurracy*) variant de 60% à 68%. Le plus haut taux (68,3%) a été observé quand les motifs ont tous la

TAB. 1 – *Résultats des tests pour le framework Multi-SPMiner avec différents paramètres. Chaque test est répété 10 fois. Les chiffres pour les relations correspondent dans l'ordre au nombre moyen de relations spatiales, spatio-temporelles et de filiation*

Nombre de nœuds du graphe	Nombre de nœuds moyen par temporalité	Nombre de relations dans les motifs	Nombre et taille des motifs dans le graphe	Taux de motifs extraits	Précision	Rappel
200	40	[5,5,2]	16 de taille 5, 13 de taille 6	0.632 ± 0.03	0.59 ± 0.03	0.68 ± 0.02
1000	40	[5,5,2]	60 de taille 5, 50 de taille 6	0.651 ± 0.02	0.63 ± 0.03	0.67 ± 0.01
1000	40	[5,5,2]	37 de taille 8, 30 de taille 10	0.664 ± 0.03	0.63 ± 0.01	0.70 ± 0.02
1000	40	[5,5,2]	90 de taille 5, 75 de taille 6	0.612 ± 0.01	0.59 ± 0.02	0.63 ± 0.01
1000	40	[5,5,2]	56 de taille 8, 45 de taille 10	0.643 ± 0.02	0.62 ± 0.01	0.64 ± 0.02
1000	40	[5,5,2]	120 de taille 5, 100 de taille 6	0.604 ± 0.02	0.59 ± 0.01	0.60 ± 0.02
1000	40	[5,5,2]	75 de taille 8, 60 de taille 10	0.631 ± 0.03	0.61 ± 0.02	0.64 ± 0.02
1000	40	[5,5,2]	150 de taille 5, 125 de taille 6	0.592 ± 0.02	0.58 ± 0.02	0.61 ± 0.01
1000	40	[5,5,2]	93 de taille 8, 75 de taille 10	0.664 ± 0.01	0.64 ± 0.02	0.67 ± 0.01
1500	40	[5,5,2]	120 de taille 5, 100 de taille 6	0.623 ± 0.01	0.62 ± 0.01	0.62 ± 0.01
1000	40	[2,2,2]	150 de taille 4, 100 de taille 6	0.634 ± 0.01	0.62 ± 0.01	0.64 ± 0.01
1000	40	[2,2,2]	188 de taille 4, 125 de taille 6	0.661 ± 0.01	0.66 ± 0.01	0.67 ± 0.01
1000	40	[5,5,2]	40 de taille 5	0.683 ± 0.04	0.66 ± 0.03	0.68 ± 0.02
2000	40	[5,5,2]	10 motifs de 4,5,6,8,10 chacun	0.647 ± 0.03	0.62 ± 0.04	0.64 ± 0.03

TAB. 2 – *Comparaison des résultats pour différentes tailles de motifs dans les expériences d'entraînement et de test.*

Nombre de nœuds	Tailles de motifs des graphes d'entraînement	Tailles de motifs des graphes de test	Ratio
1000	40 de taille 8	40 de taille 6	0.653
1000	40 de taille 10	40 de taille 6	0.642
1000	40 de taille 4	40 de taille 7	0.662
1000	40 de taille 6	40 de taille 9	0.648

même taille (avant-dernière ligne du tableau). Le taux de reconnaissance est calculé en prenant en compte le ratio des motifs identifiés, tandis que la précision et le rappel sont calculés en se basant sur l'ensemble des occurrences de ces motifs spécifiques en prennent en compte les faux positifs. Les autres paramètres testés n'ont pas affecté significativement ce taux. En comparant les expériences, nous avons observé des variations pour le taux des motifs extraits, la précision et le rappel, mais ces mesures sont restées cohérentes dans les différents scénarios. Par exemple, les expériences où la taille des graphes et les distributions de relations sont identiques ont des ratios de motifs extraits allant de 0,592 à 0,664.

Configuration 2. Les expériences menées pour évaluer la capacité de généralisation de notre modèle ont inclus des tests avec des tailles de motifs différentes entre l'apprentissage et les tests, présentés dans le tableau 2. Ces expériences ont montré que notre réseau de convolution multi-graphes (MGCN) a réussi à apprendre efficacement les relations entre les graphes et les sous-graphes, indépendamment des tailles de motifs initiales, confirmant ainsi les résultats précédents (tableau 1).

Configuration 3. Pour évaluer la robustesse de notre approche vis-à-vis de nœuds manquants dans le multigraphe, nous avons réalisé une expérience reproduisant le cinquième test du tableau 1, où des nœuds ont été supprimés de manière aléatoire. Dans la figure 3, nous montrons

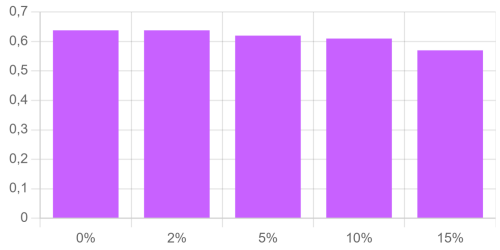


FIG. 3 – *Effet de la suppression de nœuds sur la découverte de motifs dans le graphe : Les pourcentages indiqués sous les barres représentent la proportion de nœuds supprimés, tandis que les valeurs à gauche représentent le ratio correspondant de motifs découverts.*

les résultats obtenus après avoir supprimé 2%, 5%, 10% et 15% des nœuds. L'expérience a révélé que la précision de notre modèle reste stable lorsque 2% et 5% des nœuds sont supprimés. Cependant, une légère baisse de la précision se produit avec 10% et 15% de suppression des nœuds, principalement en raison de l'impact sur la fréquence des motifs. Des instances de certains motifs ne sont plus reconnues, les rendant moins fréquents. Cependant, il est intéressant de noter que notre approche conserve une forte précision même avec la suppression de nœuds, dans une certaine limite.

Comparaison avec SPMiner. Multi-SPMiner offre un champ d'extraction de motifs plus large par rapport à SPMiner. Notre approche étend SPMiner et, lorsqu'elle est appliquée à des graphes simples, est en accord avec SPMiner, produisant des résultats comparables. Nous avons testé cela sur des graphes spatio-temporels avec uniquement des connexions spatio-temporelles (orientées et étiquettées), et des motifs de taille 6. En comparant le modèle SPMiner pré-entraîné à Multi-SPMiner entraîné à partir de zéro, SPMiner a atteint une précision de 78,7%, tandis que Multi-SPMiner a atteint 74,2%. Cette différence peut être attribuée à la taille importante des données d'entraînement de SPMiner et les contraintes liées aux propriétés d'orientation et d'étiquetage des relations pour notre modèle. En outre, Multi-SPMiner présente une complexité temporelle et spatiale comparable à celle du modèle de référence.

5 Conclusion et perspectives

Cet article présente Multi-SPMiner, une approche pour extraire des motifs dans les multigraphes en utilisant des réseaux de convolution multigraphes. Cette méthode étend les capacités de SPMiner en conservant ses principes fondamentaux tout en gérant efficacement les graphes multi-relationnels. Les résultats des expériences montrent son efficacité à découvrir des motifs fréquents dans des multigraphes uniques, suggérant son potentiel comme une version généralisée de SPMiner adaptable à divers types de graphes. Les futures recherches se concentreront sur l'amélioration de la précision de Multi-SPMiner et son application à des données réelles, notamment dans le domaine biomédical (Leborgne et al., 2021b), en utilisant des techniques d'apprentissage avancées et des architectures de graphes spécifiques.

Références

Cohn, A., B. Bennett, J. Gooday, et N. Gotts (1997). Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica 1*(3), 275–316.

Del Mondo, G., M. Rodríguez, C. Claramunt, L. Bravo, et R. Thibaud (2013). Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering 84*, 59–80.

Elseidy, M., E. Abdelhamid, S. Skiadopoulos, et P. Kalnis (2014). Grami : Frequent subgraph and pattern mining in a single large graph. *Proceedings of the VLDB Endowment 7*(7), 517–528.

Fournier-Viger, P., G. He, C. Cheng, J. Li, M. Zhou, J. Lin, et U. Yun (2020). A survey of pattern mining in dynamic graphs. *WIREs Data Mining and Knowledge Discovery 10*(6).

Ingalalli, V., D. Ienco, et P. Poncelet (2018). Mining frequent subgraphs in multigraphs. *Information Sciences 451*, 50–66.

Jiang, C., F. Coenen, et M. Zito (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review 28*(1), 75–105.

Kuramochi, M. et G. Karypis (2005). Finding frequent patterns in a large sparse graph. *Data mining and knowledge discovery 11*(3), 243–271.

Leborgne, A., M. Kirandjiska, et F. Le Ber (2021a). Random generation of a locally consistent spatio-temporal graph. In *Graph-Based Representation and Reasoning : 26th Int. Conference on Conceptual Structures, ICCS 2021*, LNCS 12879, pp. 155–169. Springer.

Leborgne, A., F. Le Ber, L. Degiorgis, L. Harsan, S. Marc-Zwecker, et V. Noblet (2021b). Analysis of brain functional connectivity by frequent pattern mining in graphs. application to the characterization of murine models. In *2021 IEEE 18th Int. Symposium on Biomedical Imaging*.

Vendrov, I., R. Kiros, S. Fidler, et R. Urtasun (2016). Order-embeddings of images and language. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ying, R., A. Wang, J. You, et J. Leskovec (2020). Frequent subgraph mining by walking in order embedding space. *The International Conference on Machine Learning (ICML)*.

Zhang, Z., P. Cui, et W. Zhu (2020). Deep learning on graphs : A survey. *IEEE Trans. on Knowledge and Data Engineering 34*(1), 249–270.

Summary

This paper introduces a novel framework, Multi-SPMiner, for frequent pattern mining in multigraphs. Building on SPMiner's foundation, Multi-SPMiner focuses on extracting frequent motifs in single multigraphs, particularly spatio-temporal graphs. It employs a two-step approach to extract high-support motifs, embedding nodes into an order space and iteratively growing motifs. The results demonstrate the effectiveness of Multi-SPMiner in identifying frequent motifs in single multigraphs, a crucial task in real-world applications. Additionally, it's tested on single connection graphs, showcasing its generality compared to SPMiner.

Index