DOSSIER DE CANDIDATURE

POUR THESE EN COTUTELLE

POUR LA RENTREE 2021

FINANCEMENT : BOURSE

Dossier complété et revêtu des signatures à transmettre impérativement pour le :
26 mars 2021 au plus tard,
A la Direction de la Recherche et Valorisation
secretariat.recherche@univ-littoral.fr

Titre de la thèse : Hybride spectral markovien par apprentissage semi-contraint et multi-échelle.

Laboratoire d'accueil ULCO : Laboratoire d'informatique Signal Image de la Côte d'Opale – ULCO LISIC UR 4491

Directeur de thèse ULCO : Emilie Poisson Caillault, MCF-HDR

Il n'y a pas de contact particulier, aussi un ordre de préférence est proposé

☐2 LIBAN – Université Libanaise (2 financements)

Pour ce dispositif, merci d'indiquer en plus :
- le nom du codirecteur étranger et le laboratoire partenaire

Pr Oussama Bazzi, Dpt of Physics and Electronics, Faculty of Science 1, Lebanese *University*

- Thématique :
  (1) La qualité de l'air
  (2) Le milieu aquatique
  (3) L'obésité, la nutrition et les activités sportives,
  (4) Les énergies propres et renouvelables
  (5) La gestion et le traitement des déchets
  (6) L'urbanisme

☐ 1 LIBAN – CNRS Libanais (4 financements)

Pour ce dispositif, merci d'indiquer en plus :
- le nom du codirecteur étranger et le laboratoire partenaire

Pr Oussama Bazzi, Dpt of Physics and Electronics, Faculty of Science 1, Lebanese *University*
- Thématique :
  (1) La qualité de l'air
  (2) Le milieu aquatique

(3) L'obésité, la nutrition et les activités sportives
(4) Les énergies propres et renouvelables
(5) La gestion et le traitement des déchets
(6) L'urbanisme

☐4 ALGERIE - Université Badji Mokhtar d'Annaba (UBMA) (2 financements)

- Thématique :

(1) La gestion et le traitement des déchets,
(2) L'aménagement littoral et portuaire,
(3) Le milieu aquatique,
(4) La surveillance et la gestion durable des Infrastructures.

☐3 MAROC – Université Mohammed V (4 financements)

- Thématique :

(1) Environnement, Milieux Littoraux Marins
(2) Sciences et technologie
(3) Santé
(4) Sciences Humaines et Sociales

*__LABORATOIRE D'ACCUEIL__

Nom du laboratoire d'accueil :LISIC, Laboratoire Informatique Signal Image de la Côte d'Opale

Nombre de HDR dans le laboratoire : 18

Nombre de thèses encadrées dans le laboratoire (rentrée 2020) : 25

Cotutelles en cours au sein du laboratoire : 12

Durée moyenne des thèses soutenues dans le laboratoire, sur la période 2015-2020 : 3,4

ENCADREMENT

Nom, Prénom du directeur de laboratoire : VEREL Sébastien

Nom, Prénom du directeur de thèse (si différent du directeur de laboratoire) : POISSON CAILLAULT Emilie

Nombre de doctorats en préparation sous la direction du directeur de thèse : 1

Avis détaillé du directeur de thèse :

Ce projet de recherche s'intègre dans une dynamique forte de travaux liés à la classification et modélisation de séries temporelles, développés pour l'aide à l'interprétation fine de la dynamique d'un processus tel que les pollutions algales, fluviales. D'un point de vue applicatif, ces outils permettent d'analyser le fonctionnement d'un milieu naturel et aident à la mise en place de nouvelle stratégie d'échantillonnage pour gérer au mieux et améliorer la qualité de l'eau.
Ce sujet bénéficie d'une part d'une expérience de recherche effective conjointe attestée par une délégation de 2014 à 2016 à IFREMER (institut Français de Recherche sur la Mer), plusieurs soutenances de thèse (Wacquet 2013, Rousseeuw 2014, PHAN 2019, Grassi 2020) et de ma propre HDR (7 février 2020). D'autre part, il est appuyé par une valorisation logicielle (uHMM, DTWBI, DTWUMI, RclusTool) ouverte à la communauté scientifique mais aussi des interfaces intégrées permettant aux utilisateurs finaux de traiter leurs données. Ceci a été notamment récompensé par l'attribution du premier Prix Partenariat du CEREMA 2019 pour l'étude de la réserve naturelle du Marais de l'Isle-Saint-Quentinn à partir d'un hybride markovien non supervisé.
Ces compétences seront renforcées par des collaborations actives avec Messieurs Oussama Bazi (Pr, Univ. Liban), Alain Lefebvre (Dir . LER-BL IFREMER, HDR) et André Bigand (MCF, HDR, ULCO-LISIC) pour leurs expertises respectives en traitement du signal, eutrophisation des milieux marins, gestion de l'incertitude assureront une couverture optimale des domaines scientifiques associés au sujet mais aussi des projets phares tel JERICO S3 ou des actions de la SFR MER, dont je suis membre au conseil scientifique.

Signature du directeur de thèse

E. Poisson Caillault.

Avis détaillé du directeur de laboratoire :

Le développement des activités du laboratoire dans le domaine de l'apprentissage automatique (machine learning) correspond à une priorité de notre unité de recherche. Cette thèse s'inscrit dans les travaux de l'équipe IMAP qui concerne l'apprentissage non-supervisé et vient renforcer cette priorité du laboratoire. L'application cible et support des travaux de thèse fait suite à des travaux déjà initiés autour d'un réseau de collaborations solidement établies, et s'inscrivant là encore dans une priorité du laboratoire (SFR Campus de la Mer).
Pour toutes ces raisons, j'émets un avis favorable à son financement.

Signature du directeur de laboratoire

PROJET DE THESE

Intitulé du projet de thèse :

Hybride spectral markovien par apprentissage semi-contraint et multi-échelle.

Applications à la caractérisation des événements nuisibles d'un milieu naturel et notamment la qualité de l'eau.

(Hybrid Spectral Clustering -Hidden Markov Model by constraint and multi-scale learning.)

Domaine scientifique : Traitement du Signal / Informatique

Résumé (1/2 page maxi.) :

L'objectif de thèse est de faire émerger des méthodes capables d'intégrer l'ensemble des informations disponibles des processus physico-biologiques, acquises à des fréquences hebdomadaires, mensuelles ou annuelles directement dans l'apprentissage/construction des systèmes de détection et prédiction actuels qui travaillent sur des échelles beaucoup plus fine (de la seconde à 20 minutes). La connaissance actuelle des processus est souvent partielle et incomplète aussi bien temporellement que spatialement, notamment à cause d'opérations de surveillance dites « coup-de-poing » ou de campagnes en mer sur des périodes données.

Les approches de clustering spectral contraint, clustering multi-échelle et modélisation markovienne ont démontré leur intérêt comme aide à l'interprétation des événements d'un milieu naturel. Il convient maintenant d'apporter une interprétation intégrée tant sur la connaissance de la phénologie des processus que sur le niveau d'interprétation souhaitée (approche globale à fine).

Les travaux de recherche liés à cette thèse impliqueront un consortium de deux membres HDR du LISIC (E. Caillault et A. Bigand), le Professeur Oussama Bazzi (Université Libanaise) et le chercheur expert Alain Lefebvre (Dir. LER-BL/IFREMER, HDR).

Mots clés : séries temporelles, incertitude, apprentissage par contraintes, apprentissage multi-échelle, classification semi-supervisée, intervalle de confiance 2D, qualité de l'eau.

Projet de thèse (5 pages maxi.)

Le sujet de recherche choisi et son contexte scientifique

Dans le domaine de l'écologie numérique, les méthodes de classification présentent un grand intérêt pour synthétiser l'information, comprendre la structure des données et ensuite extraire le maximum d'informations viables. Ces méthodes sont utilisées pour améliorer la connaissance et obtenir des recommandations pratiques pour la gestion de l'environnement. Dans le contexte d'amélioration de la qualité des eaux maritimes et de démarche écologique (MSFD, 2008) ou d' « état eutrophique » (OSPAR, 2002), la connaissance de la dynamique des écosystèmes à une échelle temporelle fine est souvent incomplète et nécessite une approche non-supervisée (Ferreira, 2011). Dans ce type d'approche non-supervisée [Jain 2010], le système de classification extrait l'information à partir de données brutes (par ex. les paramètres physico-chimiques, les capteurs multiples) et détecte les états environnementaux particuliers (« clustering » spectral de données de cytomètres [CSC, 2003 ; ICCE, 2016], modèle de Markov caché non-supervisé uHMM [uHMM, 2015]) Dans les travaux de recherche récents, deux types d'information a priori se distinguent. La première connaissance est relative à la dynamique de la fluorescence et la seconde est relative à la connaissance de la biomasse du plancton (phyto- ou zoo-planctons) acquise à partir de nouvelles technologies sur de courtes périodes. Il est cependant approprié de prendre en compte cette information supplémentaire dans le processus de classification et de modélisation, même si l'échelle des données ou leur fréquence est faible.

De nombreux hybrides profonds markoviens ont émergé ces cinq dernières années, ainsi que des approches multi-échelles combinant l'extraction d'information à plusieurs niveaux avec une unique étape de décision. L'approche multi-échelle est réduite à la fusion des couches de réseaux de neurones (souvent à convolutions) en entrée d'un perceptron qui fournira les probabilités d'observation d'être dans tel état du HMM (Thomas, PAA 2015; Wetteland, MIDL 2019). L'approche multi-échelle de la décision et donc de l'interprétation est alors perdue. Des structures hiérarchiques de HMM (HHMM) ont émergé à partir des années 2000 et l'optimisation de leur topologie a été testé pour les modèles de langage (Wakabayashi, 2010) mais nécessite des corpus d'apprentissage large et des a priori forts. L'hybridation d'un HMM via une classification spectrale profonde semble une voie prometteuse pour aboutir à un HHMM non supervisé sans *a priori* nécessaire à la définition

de sa topologie et permettre une intégration de connaissances faibles (par opposition à l'apprentissage supervisé avec connaissance de labels).

Le qualité de l'eau au Liban est fortement dépendante de son emplacement dans le bassin-est de la Méditerranée et de sa topographie montagneuse et d'autre part des changements climatiques. Assurer son monitoring requiert alors d'intégrer l'ensemble des informations disponibles. L'écosystème associé aux côtes libanaises pourra être comparé à celui du bassin Nord de la Méditerranée (données satellitaires, bouées MAREL Mesurho).

L'état du sujet dans le laboratoire et l'équipe d'accueil

Ce projet de thèse s'inscrit dans une thématique forte de l'équipe IMAP, l'apprentissage automatique et un projet phare de l'université du Littoral : l'Environnement au travers de collaborations fortes au sein de la Structure Fédérative de Recherche MER, d'un Contrat Plan Etat Région MARCO et la participation au projet H2020 JERICO-S3, projet d'intégration de la recherche côtière en Europe.

Les trois axes sous-jacent à la thèse proposée sont la classification spectrale semi-supervisée sur des données multidimensionnelles, la classification spectrale multi-échelle et la modélisation markovienne par apprentissage non supervisé à partir de séries temporelles multidimensionnelles. Ces trois volets ont été amorcés et validés par plusieurs thèses décrites ci-dessous et des publications et journaux reconnus (JSTARS, IGARSS, PRL, CI, MEPS).

Dans ce contexte, 4 thèses ont été soutenues en 3 ans. La première thèse, soutenue en décembre 2011 et co-encadrée par E. Poisson Caillault (LISIC) a porté sur la classification spectrale contrainte [via le formalisme de Wagstaff et Cardie (Wagstaff, 2000)], appliquée à la classification de cellules phytoplanctoniques à partir de données cytométriques. La seconde thèse, soutenue le 11 décembre 2014 et co-encadrée par E. Poisson Caillault (LISIC) et A. Lefebvre (IFREMER LER-BL), a permis d'établir un Modèle de Markov Caché (MMC/HMM) par apprentissage non supervisé et sans connaissance a priori. Ce modèle a permis de construire un système automatique d'estimation d'états environnementaux caractéristiques à partir des mesures à haute résolution temporelle avec les aléas engendrés de données manquantes ou aberrantes. La troisième thèse, soutenue en 2018 et co-encadrée par E. Poisson Caillault (LISIC) et A. Bigand (LISIC) concernait l'imputation de séquences manquantes dans les séries temporelles afin d'améliorer les processus de classification et prédiction d'événements futurs. La quatrième thèse soutenue en novembre

2020 concerne le clustering et la classification multi-échelle multi-sources (satellitaires, bouées MAREL, Réseau SRN).

Dans [GDR Phytocox 2016, Ocean 2017], nous avons proposé une approche de segmentation EM pour des séries temporelles de Chlorophylle-a afin d'améliorer la caractérisation de la dynamique de Chl-a sur le site de Gravelines. Cette approche permet d'identifier les événements environnementaux ainsi que la date de leur début et de leur fin. Un ensemble de données collectées durant un événement peut être associé au même groupe. De même, des données non-incluses dans cet événement appartiendront à deux groupes différents.

Les travaux (Ali Rizik, rapport de stage recherche, master 2 de l'Université Libanaise au LISIC en 2016) ont montré la faisabilité et l'intérêt d'intégrer une connaissance temporelle a priori (connaissance relative au « bloom » de fluorescence) en utilisant des contraintes de type « Must-Link ». Ces contraintes sont directement intégrées dans la matrice de similarité du « clustering » spectral.

Cette thèse est donc la continuité logique des travaux de l'équipe et permettra de répondre aux questions suivantes :

- Est-il possible d'introduire dans la modélisation markovienne une structure et une dynamique semi-contrainte et multi-échelle en insérant une information temporelle ?
- La formalisation des critères de contraintes (appelés usuellement dans la littérature et introduit par Wagstaff, Must Link et Can Not Link) peut être elle insérer facilement dans le critère d'optimisation du HHMM ? La confiance dans la connaissance a priori ajoutée sera modélisée par un intervalle de confiance flou « 2D », [Bigand 2010, 2016]
- Comment détecter un événement non appris et réapprendre le modèle ?

Le programme et l'échéancier de travail

Le travail de thèse devrait se décomposer de la façon suivante :

Année 1 : Bibliographie, récolte et exploration des données disponibles pour le monitoring de la qualité de l'eau. Prise en main des outils existants et de la littérature : uHMM (SC-HMM), Modèle de Markov Caché par apprentissage non supervisé, MSC - classification spectrale Multi-échelle, cSC : classification spectrale contrainte.

Année 2 : Développement et simulation d'un système hybride spectral markovien par apprentissage semi-contraint et multi-échelle. Introduction et définition du rejet et de l'apprentissage incrémental. Valorisation et communication.

Année 3 : Simulation de l'apprentissage dynamique et développement d'indicateurs. Rédaction du manuscrit. Valorisation et communication.

Les retombées scientifiques et économiques attendues

L'apprentissage d'un modèle de Markov Caché par apprentissage non supervisé guidé ouvre une porte considérable pour traiter des applications réelles où la taille des séries temporelles collectées est telle qu'il n'est plus possible de demander à ces échelles, vu la multiplicité et variété des capteurs/ mesures, de demander un étiquetage humain de chaque événement. Un apprentissage supervisé est donc inimaginable et source d'erreurs importantes. L'apprentissage semi-supervisé reste la piste la plus cohérente puisque nous pouvons disposer de quelques connaissances a priori et intégrer par classification spectrale la géométrie des données.

D'un point de vue applicatif, la prise en compte des données des données acquises par les différents dispositifs de collecte de données (basse et haute fréquences temporelle et/ou spatiale) en milieu marin permet d'envisager une approche multi-paramètres et multi-échelles indispensable pour une approche écosystémique telle que recommandée dans la mise en œuvre de la Directive Cadre Stratégique pour le Milieu Marin. Régionalement, la mise à disposition d'un outil numérique capable d'apprendre et de détecter les événements rares et récurrents en temps réel peut s'avérer utile pour les besoins de surveillance des Parcs Marins et Réserves naturelles.

Dans le futur, ce sujet a aussi des perspectives d'application dans les réseaux capteurs et IoT afin de compenser les pertes de données lors de la transmission quel que ce soit l'application lié à l'environnement naturel.

Les collaborations prévues et une liste de 10 publications maximum portant directement sur le sujet

Ces travaux seront effectués en étroite collaboration avec les partenaires de la SFR Mer et du CPER MARCO , des partenaires du projet H2020 JERICO-S3 dont notamment Ifremer.

1- (CSC, 2003) Guillaume Wacquet; Emilie Poisson Caillault; Denis Hamad; Pierre-Alexandre Hébert. Constrained Spectral Embedding for K-Way Data Clustering. Pattern Recognition Letters, Available online 19 February 2013, ISSN 0167-8655, doi:10.1016/j.patrec.2013.02.003

2- (uHMM, 2015) Kévin Rousseeuw, Emilie Poisson-Caillault, Alain Lefebvre, and Denis Hamad. Hybrid Hidden Markov Model for Marine Environment Monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, DOI 10.1109/JSTARS.2014.2341219. Volume 9 (issue 1), published in 2015.

3- (MSC, 2019) K. Grassi, E. Poisson Caillault and A. Lefebvre, "Multilevel Spectral Clustering for extreme event characterization," *OCEANS 2019 - Marseille*, Marseille, France, 2019, pp. 1-7. doi: 10.1109/OCEANSE.2019.8867261

4- (HHMM, Wakabayashi 2010) Wakabayashi K., Miura T. (2010) Topology Estimation of Hierarchical Hidden Markov Models for Language Models. In: Natural Language Processing and Information Systems. NLDB 2010. Lecture Notes in Computer Science, vol 6177. Springer, Berlin, Heidelberg

5- (MEPS, 2019) Lefebvre A, Poisson-Caillault E (2019) High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering. Mar Ecol Prog Ser 608:73-92. https://doi.org/10.3354/meps12781

6-(Wagstaff, 2000) WAGSTAFF , K. AND C ARDIE , C. 2000. Clustering with instance-level constraints. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, Palo Alto, CA, 1103–1110.

7- (Ocean, 2017) Emilie Poisson Caillault and A. Lefebvre. Towards Chl-a Bloom Understanding by EM-based Unsupervised Event Detection. Ocean conference in abeerden. June, 2017.

8- (Bigand 2016) André Bigand, O. Colot : « Membership functions construction for interval-valued fuzzy sets with application to Gaussian noise reduction “, Revue FSS (Fuzzy Sets and Systems), 286, pp.66-85, 2016.

Citer dans le texte :

(MFSD, 2008). DIRECTIVE 2008/56/EC OF THE EUROPEAN PARLIAMENT and OF THE COUNCIL of 17 June 2008 establishing a framework for community action in the field of marine environmental policy. Msfd : Marine strategy framework directive. Official journal of the european union, 2008.

(Ferreira, 2011). João G. Ferreira, Jesper H. Andersen, Angel Borja et al.. Overview of eutrophication indicators to assess environmental status within the european marine strategy framework directive. Estuarine, Coastal and Shelf Science, 93(2):117 – 131, 2011.

(Jain, 2010). A. K. Jain, «Data Clustering: 50 Years Beyond K-Means" , Pattern Recognition Letters, Vol. 31, No. 8, pp. 651-666, 2010.