DOSSIER DE CANDIDATURE A UNE ALLOCATION DE RECHERCHE POUR LA RENTREE 2016

Titre de la thèse : Modèle de Markov Caché par Apprentissage semi-contraint. Application à la caractérisation des efflorescences phytoplanctoniques à partir de données hautes fréquences.

Laboratoire d'accueil ULCO : ULCO/LISIC, Laboratoire d'Informatique Signal Image de la Côte d'Opale.

Web : http://www-lisic.univ-littoral.fr/

Directeur de thèse ULCO : André BIGAND

Partenaire étrangers si identifié (noms de la structure de recherche et du codirecteur étranger) : Y.Mohanna et O.Bazzi, Professeurs, Ecole Doctorale de l'Université Libanaise (UL), LIBAN

-Thématique : milieux aquatiques

***LABORATOIRE D'ACCUEIL**

Nom du laboratoire d'accueil : **LISIC**
Nombre de HDR dans le laboratoire : **12**
Nombre de thèses encadrées dans le laboratoire (rentrée 2014) : **26 (dont 15 cotutelles)**
Durée moyenne des thèses soutenues dans le laboratoire, sur la période 2010-2014 : **39 mois**

ENCADREMENT

Nom, Prénom du directeur de laboratoire : **Renaud Christophe**

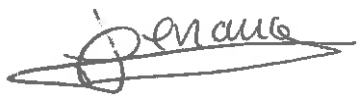
Nom, Prénom du directeur de thèse : **Bigand André**
Nombre de doctorats en préparation sous la direction du directeur de thèse : **2**

Avis détaillé du directeur de thèse :
Ce projet de recherche s'intègre dans une dynamique forte de collaboration avec l'Université Libanaise (UL). Il bénéficie d'une expérience de recherche effective avec l'UL attestée par une soutenance de thèse en co-tutelle en 2013 (H.Hijazi) et d'une thèse en cours (A.Darwich).
Il bénéficie d'autre part de l'expertise de l'équipe IMAP dans le domaine (thèse de M.Rousseeuw, 2014) et d'un support scientifique et logistique fort de l'Ifremer. Je suis donc très motivé pour assumer la direction d'un étudiant Libanais sur le sujet décrit dans ce document.

Signature du directeur de thèse
cf. version papier

Avis détaillé du directeur de laboratoire :

Le projet de thèse proposé se situe dans des domaines scientifiques pour lesquels les encadrants, tant du côté ULCO que du côté libanais, disposent de compétences élevées. Le projet cadre également parfaitement avec les objectifs prioritaires de l'ULCO (campus de la mer) et l'un des objectifs fixés dans le cadre de ce type de demande (milieux aquatiques) et pourra bénéficier de l'appui d'un partenaire avec lequel le laboratoire développe des recherches depuis plusieurs années. Enfin, ce sujet se situe dans les problématiques scientifiques développées au sein de l'équipe IMAP du laboratoire qui a été bien évaluée par l'AERES en 2014. L'avis quant au financement de ce projet est donc très favorable.



Signature du directeur de laboratoire

PROJET DE THESE

Intitulé du projet de thèse : **Modèle de Markov Caché par Apprentissage semi-contraint. Application à la caractérisation des efflorescences phytoplanctoniques à partir de données hautes fréquences.**

Domaine scientifique : **Traitement du signal, classification de données, données environnementales, écologie numérique.**

Résumé (1/2 page maxi.) :

Rousseeuw et al. [1] ont montré théoriquement et expérimentalement qu'il est possible de générer un Modèle de Markov Caché (MMC) par apprentissage non supervisé. L'utilisation de la théorie associée à la classification spectrale a alors permis de mettre en avant une structure et caractérisation du MMC cohérente vis-à-vis des observations sans aucune intervention de paramétrage, sans aucun *a priori* sur la structure. Cette technique a été appliquée pour la première fois pour modéliser la dynamique des efflorescences phytoplanctoniques en Manche orientale et détecter des états particuliers à partir d'une base de données dite « haute résolution » dans le domaine de l'observation marine côtière (fréquence : 20 min.).

L'objectif de cette thèse est d'une part d'évaluer l'applicabilité des techniques d'apprentissage semi-supervisé, afin de déterminer un MMC fidèle à la connaissance existante à la fois en terme de structure et de temporalité et d'autre part proposer un système robuste à des événements/états extrêmes non appris. Ces états nécessitent la définition d'un rejet du système existant [1] et la mise en place d'un apprentissage dynamique.

Une telle technique pourrait ainsi permettre de traiter un nombre important de données issues de stations instrumentés « haute fréquence » (fréquence infra-horaire) tout en insérant les connaissances acquises depuis des décennies via les approches conventionnelles dites « à basse résolution » (prélèvements mensuels ou bimensuels, rarement hebdomadaires). En Manche orientale, par exemple, il s'agira d'intégrer les données issues de la station MAREL Carnot (Station de mesures instrumentées autonomes, multi paramètres) avec les données des réseaux d'observation comme le REPHY/SRN (Réseaux Phytoplancton & Phycotoxines , et Suivi Régional des Nutriments) afin de mieux comprendre la dynamique phytoplanctonique dans le contexte du développement d'algues nuisibles, en insistant particulièrement sur les effets pressions/impacts (e.g. effets directs et/ou indirects des apports de nutriments) afin de contribuer aux enjeux des directives ou conventions de mers régionales.

Cette thèse s'effectuera en collaboration avec l'UL et le laboratoire IFREMER, LER/Boulogne-sur-Mer, qui mettra à disposition les données, la plateforme logicielle et son expertise en écologie marine.

[1] Rousseeuw, K., Poisson-Caillault, E., Lefebvre, A. and Hamad, D. "Hybrid Hidden Markov Model for Marine Environment Monitoring", in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, doi 10.1109/JSTARS.2014.2341219. 20 août 2014.

http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6880782

Projet de thèse (5 pages maxi.) :

Modèle de Markov Caché par Apprentissage semi-contraint. Application à la caractérisation des efflorescences phytoplanctoniques à partir de données hautes fréquences.

Mots-clés : Modèle de Markov caché, apprentissage dynamique, apprentissage semi-supervisé, classification spectrale par contrainte, efflorescences phytoplanctoniques.

Cette thèse a pour objet de proposer un apprentissage semi-supervisé d'une modélisation markovienne de phénomènes environnementaux à partir de données multidimensionnelles haute fréquence. Les aspects apprentissage automatique et apprentissage semi-supervisé sont des thématiques clés de l'Université Libanaise (UL, équipe des professeurs Y. Mohanna et O. Bazzi, Ecole Doctorale de l'Université Libanaise, LIBAN) et de l'université du Littoral Côte d'Opale (Laboratoire LISIC, équipe IMAP) dont les collaborations ont déjà été amorcées depuis 2011. Les collaborations du LISIC avec Ifremer (LER Boulogne-sur-Mer) existantes depuis 2008 permettront d'asseoir l'interprétation biologique des données et seront étendues avec l'UL.

Le sujet de recherche choisi et son contexte scientifique

Les milieux aquatiques au Liban ont été très affectés par son histoire (guerre des années 1980, guerre de 2006) et par une pollution importante. Cette longue période a été très destructrice pour la qualité des eaux fluviales et maritimes. Dans ce contexte, un réseau de surveillance de la qualité des eaux est primordial.
La prise de conscience générale des problèmes d'environnement, notamment au niveau du littoral (Français et Libanais), conduit à renforcer l'observation et la surveillance qui s'y exerce. Par l'expérience acquise depuis de nombreuses années dans l'exploitation des réseaux de surveillance de l'environnement, l'Ifremer a mis en évidence le besoin de développer des systèmes de surveillance automatisée de l'environnement et des effets directs et indirects des activités humaines sur le milieu marin. Les développements technologiques concernant les capteurs physico-chimiques permettent la mise en œuvre de réseaux de stations instrumentées autonomes, effectuant des mesures à fréquence dite « haute résolution – haute fréquence (HF)» dans le domaine de l'observation marine côtière (fréquence d'échantillonnage de l'ordre de quelques minutes à quelques heures) et rapidement disponibles pour les utilisateurs (portail web).

L'ensemble des mesures acquises maintenant, depuis 2004 pour citer celles de la station MAREL Carnot de Boulogne-Sur-Mer, permettent aujourd'hui d'appréhender des thématiques de recherche, comme la détection et la prédiction des efflorescences phytoplanctoniques via des ***systèmes d'apprentissage automatique***. Un nombre important de travaux ont été dédié à la prédiction d'algues toxiques dans un contexte de classification à deux évènements (présence ou absences d'algues toxiques). Très récemment, des travaux de ***modélisation markovienne par apprentissage non supervisé*** ont montré qu'il était possible de détecter automatiquement puis caractériser plus finement les états environnementaux caractéristiques de la dynamique planctonique et des paramètres physico-chimiques, facteurs de contrôle ou reflet des effets directs et indirects des développements de phytoplancton, mais aussi des événements liés à des activités humaines, aux conditions hydrodynamique ou météorologique [1, 2]. Notamment, le système markovien, hybridé via une ***classification spectrale*** pour générer automatiquement la structure du Modèle de Markov Caché (MMC), développé dans [1] a permis de mettre en évidence une succession de signatures caractéristiques des états cohérente vis-à-vis de l'expertise écologique. Afin de caractériser précisément ces états et prédire les états d'alertes d'espèces nuisibles, il convient d'analyser la biomasse et la composition phytoplanctonique à la même échelle que les observations des paramètres physico-chimiques. Certes, un certain nombre de programmes d'échantillonnage ont été mis en place afin d'observer et de surveiller l'état du milieu marin permettant ainsi de caractériser la biomasse et/ou

composition phytoplanctonique mais avec des suivis à fréquence mensuelle ou bimensuelle, plus rarement à fréquence hebdomadaire (par convention appelée approches à basse fréquence ou basse résolution et les méthodes de mesures et d'analyses associées sont dites conventionnelles). Dans ces échelles de résolution, on peut citer le réseau national de surveillance du phytoplancton et des phycotoxines, nommé REPHY créé en 1984 avec des prélèvements bimensuels. Depuis 1992, une extension vers le large de ce réseau REPHY est réalisée régulièrement via le réseau Suivi Régional des Nutriments (SRN) [3,4].
La prise en compte des connaissances même à des échelles différentes permettrait d'améliorer le système de modélisation et prédiction des états. Récemment, ***l'apprentissage semi-supervisé*** a reçu une attention toute particulière dans les approches discriminantes tels que les classifieurs spectraux ou machines à vecteurs supports. Cet apprentissage semi-supervisé permet de faire intervenir des connaissances a priori dans un processus de ***décision à partir de contraintes sur les données*** (association de points appartenant à la même classe/état ou exclusion) ou de ***labellisation réduite*** (quelques états sont connus). Un ensemble de travaux [5,6] ont montré la puissance de ces approches pour réaliser des partitionnements entre données, cohérents à la fois vis-à-vis de la structure géométrique des données et la connaissance existante. Il est alors intéressant d'étendre cela à la modélisation de données spatio-temporelles.

L'objectif de cette thèse est ainsi d'une part d'évaluer l'applicabilité des techniques d'apprentissage semi-supervisé, notamment de ***classification spectrale contrainte*** afin de déterminer un MMC fidèle à la connaissance existante à la fois en terme de ***structure et de temporalité*** et, d'autre part, de proposer un système robuste à des événements/états extrêmes ou non appris. Ces états nécessitent la définition d'un ***rejet*** lors de la prédiction d'une nouvelle donnée éloignée des observations/données existantes et la mise en place d'un ***apprentissage dynamique*** de ces nouveaux états.

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

Ce projet repose sur une longue coopération avec l'Université Libanaise (UL) au travers d'encadrement de stages de Master et de cotutelles de thèses. Cette collaboration a d'abord été initiée par D. Hamad à Tripoli (Faculté de Génie 1) et à Beyrouth, (facultés des sciences, branches 2 et 3), puis poursuivie par A. Bigand à Beyrouth (UL 1) à partir de 2011. A partir de cette date, la collaboration avec l'UL s'est d'abord déroulée sous forme d'un encadrement de thèse en co-tutelle (Hala Hijazi, co-dirigée par le Prof. O.Bazzi, UL, soutenance effectuée le 19/12/13). En parallèle, le Prof. O.Bazzi et A . Bigand avons mis en place une collaboration pédagogique qui a débouché sur la mise en place d'un co-diplôme entre les master INS3I (ULCO) et STIP (UL). La première étape de cette collaboration repose sur l'échange d'étudiants pour le stage de master 2 et la définition d'un cours à l'UL (Pattern Recognition and Machine Learning). La deuxième étape de cette collaboration est constituée par la montée en puissance des objectifs précédents (stages de master 2, échange de cours) et la poursuite de nos travaux de recherche communs sous forme d'une nouvelle thèse (A.Darwich).

D'autre part, ce projet s'inscrit aussi dans le **projet ARCUS 2**, déposé à l'ULNF avec le Maroc, la Palestine et le Liban. H. Hijazi [8,9] a mis au point des outils d'interprétation et de visualisation augmentée de données, par l'intermédiaire de méthodes de réduction de la dimension pour l'analyse exploratoire de données multidimensionnelles. Ces méthodes sont étendues à l'apprentissage semi-supervisé (thèse de co-tutelle de M. A. Darwich démarrée en 2014) et devraient être appliquées au modèle de Markov Caché par Apprentissage semi-contraint sans problème par le biais de ce projet.
Le travail se fera donc en étroite collaboration avec l'UL. Il s'appuiera sur une collaboration active avec des collègues de l'Université Libanaise par une application des méthodes obtenues par le doctorant aux problèmes de ***pollution marine au Liban*** et des publications communes.

Ce projet de thèse s'inscrit dans une thématique forte de l'équipe IMAP, ***l'apprentissage automatique*** et un projet phare de l'université : l'***Environnement*** au travers de collaborations fortes avec l'IFREMER et le LOG depuis 2008 sur l'étude et le suivi de communautés phytoplanctoniques pour la surveillance de l'écosystème marin, un projet INTERREG IVa 2 mers DYMAPHY et une

collaboration ULCO/LISIC-Ifremer/LER - Agence de l'Eau Artois Picardie. Ce projet s'inscrit dans la continuité de la dynamique de recherche engagée dans le cadre des activités du **Groupement d'Intérêt Scientifique (GIS) « Campus International de la Mer et de l'Environnement Littoral »** regroupant l'ensemble des acteurs de l'enseignement supérieur et de la recherche, des collectivités territoriales et du monde socio-économique dont les recherches sont relatives à la mer et à l'environnement littoral sur le territoire Manche-Mer du Nord.

Les deux axes sous-jacent à la thèse proposée sont la classification spectrale semi-supervisée sur des données multi dimension et la modélisation markovienne par apprentissage non supervisé à partir de séries temporelles multi dimension. Ces deux volets ont été amorcés et validés par la thèse de Hala Hijazi au Liban et les thèses décrites ci-dessous au LISIC et ont fait l'objet d'une valorisation via des communications, l'édition de publications dans des revues internationales à comité de lecture [1,2,5,6,8,9].

Dans ce contexte, 2 thèses à l'université du Littoral (LISIC) 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 appliquée à la classification de cellules phytoplanctoniques à partir de données cytométriques. Monsieur Wacquet est actuellement en post-doctorat à l'université de Mons. 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) par apprentissage non supervisé. Ce modèle a été 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.

Cette thèse s'inscrit donc dans la continuité des travaux des équipes et permettra de répondre aux questions suivantes :
- Est-il possible d'introduire dans la modélisation markovienne une structure et une dynamique semi-contrainte ?
- La formalisation des critères de contraintes (appelés usuellement dans la littérature et introduit par Wagstaff en 2005 « Must Link » et « Can Not Link » i.e. deux points doivent appartenir au même ensemble ou ne doivent pas) peut être elle insérer facilement dans le critère d'optimisation du MMC ?
- Comment détecter un évènement non appris et réapprendre le modèle ?
- L'information Basse fréquence introduite dans le système existant MMC-NS permet-elle de confirmer l'expertise humaine ?

Cette thèse bénéficierait également de la dynamique générée par des projets complémentaires, comme le **CPER-MARCO** « Recherches marines et littorales en Côte d'Opale : des milieux aux ressources, aux usages et à la qualité des produits aquatiques qui se veut être un projet structurant multi-laboratoires, multi-organismes associant la mise en place d'instruments et d'outils (enquêtes, indicateurs) pour une approche globale de l'étude du milieu marin, de la ressource et de la qualité des produits aquatiques » et le projet **H2020 JERICO-Next** (New European eXpertise for coastal observatories, http://www.jerico-fp7.eu/).

Le programme et l'échéancier de travail

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

- Année 1 : Bibliographie, analyses des données et prise en main du modèle existant de Markov Caché par apprentissage non supervisé. Adaptation de l'algorithme de construction par apprentissage semi-supervisé.
- Année 2 : Développement et simulation du modèle sur les données des stations instrumentées. Introduction et définition du rejet et de l'apprentissage dynamique. Valorisation et communication.

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

Le candidat mènera alternativement ses travaux de recherche dans chacun des établissements partenaires. Celui-ci effectuera des séjours de recherche auprès de chacun des établissements. La durée des périodes de séjour sera annuellement : 6 mois à l'UL et 6 mois à l'ULCO (voir le tableau ci-dessous), en accord avec les deux établissements, pendant l'application de la présente convention :

	ULCO	**Université Libanaise**
2016-2017	1 Janvier 2017 – 30 Juin 2017	1 Septembre 2016 – 30 Décembre 2016
2017-2018	1 Janvier 2018 – 30 Juin 2018	1 Septembre 2017 – 30 Décembre 2017
2018-2019	1 Janvier 2019 – 30 Juin 2019	1 Septembre 2018 – 30 Décembre 2018

Les retombées scientifiques et économiques attendues

L'apprentissage d'un modèle de Markov Caché par apprentissage non supervisé ouvre une porte considérable pour traiter des applications réelles où la taille des séries temporelles collectées est tel qu'il n'est plus possible de demander à ces échelles, vu la multiplicité et variété des capteurs/ mesures, un étiquetage à dire d'expert 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.

Ces travaux contribueront inévitablement à l'amélioration des connaissances en ce qui concerne la dynamique des efflorescences du phytoplancton et de leurs effets directs et indirects sur l'environnement, comme sur la santé humaine (cas des efflorescences d'espèces productrices de phycotoxines). La définition d'états caractéristiques d'une dynamique ou d'événements extrêmes permettra de définir des stratégies d'échantillonnage optimisées pour le besoin des programmes d'observation et de surveillance de la qualité de l'environnement marin. La prédiction de ces états permettra de développer des systèmes d'alertes permettant ainsi aux managers de l'environnement de proposer des mesures, des réponses adaptées. Par ailleurs, la prise en compte 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 par la Politique Marine de l'Union Européenne au travers de la mise en œuvre de la DCSMM. Régionalement, le déploiement de ce genre de système de mesures automatisées à haute fréquence peut s'avérer utile pour les besoins de surveillance des Parcs Marins.

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

Ces travaux seront effectués en étroite collaboration avec l'UL et les partenaires du Campus de la Mer/CPER MARCO, Université de Lille I, Université de l'Université Côte d'Opale, UMR CNRS 8187 LOG, Ifremer LER-BL. Par ailleurs, l'ensemble des laboratoires Environnement & Ressources de l'Ifremer qui déploient des systèmes qui vont de la sonde multiparamètres de base aux systèmes plus évolués de mesures HF (marel estran, bouée smatch …) pour les besoins des réseaux de surveillance ou pour des études et recherches seront associés à la démarche afin d'envisager le

transfert des méthodologies développées dans cette thèse à leurs problématiques (efflorescence, eutrophisation, observatoire conchylicole …).

Les actions menées par le LER/BL en conformité avec son **thème de recherche « comprendre la dynamique planctonique et développer des outils d'analyses automatisées »**, le développement de son expertise sur les sujets liés (hydrologie, phytoplancton, zooplancton) et les collaborations mises en place aux échelles régionale, nationale et internationale permettront au laboratoire de contribuer activement à ce projet (plateformes instrumentées, base de données, expertises en écologie numérique et marine,…).

Bibliographie

[1] Rousseeuw, K., **Poisson-Caillault, E., Lefebvre, A.** and Hamad, D. "Hybrid Hidden Markov Model for Marine Environment Monitoring", in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, doi 10.1109/JSTARS.2014.2341219. 20 août 2014. (IF: 2.86) **http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6880782**

[2] Rousseeuw, K., **Poisson-Caillault, E., Lefebvre, A.** and Hamad, D. "Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling.", proceedings of *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2013, Melbourne, Australia, 21-26 July 2013. pp 3692-3965.* http://dx.doi.org/10.1109/IGARSS.2013.6723700

[3] Hernández Fariñas T., Soudant D., Barillé L., Belin C., **Lefèbvre A.**, Bacher C, "Temporal changes in the phytoplankton community along the French coast of the eastern English Channel and the southern Bight of the North Sea". *ICES Journal of Marine Science* 71 (4), pp. 821-833. (IF: 2.525), 2014.

[4] **Lefebvre A.** , Guiselin N., Barbet F., Artigas L. F., "Long-term hydrological and phytoplankton monitoring (1992-2007) of three potentially eutrophicated systems in the eastern English Channel and the southern bight of the North Sea. *ICES Journal of Marine Science*, 68(10), pp. 2029-2043, 2011

[5] Wacquet, G., **Poisson Caillault E.**, Hébert P.A., "Semi-supervised K-Way SpectralClustering with Determination of Clusters », in *Computational intelligence, Series Studies in Computational Intelligence*, Springer, vol 465, pp. 317-330, isbn={978-3-642-35637}, 2013.

[6] Wacquet, G., **Poisson Caillault E.**, Hamad, D., Hébert P.A., 'Constrained Spectral Embedding for K-Way Data Clustering », in Pattern Recognition Letters, d*oi:10.1016/j.patrec.2013.02.003*, v.34 n.9, p.1009-1017, July, 2013 *2013.*

[7] **André Bigand**, Olivier Colot, Fuzzy filter based on interval-valued fuzzy sets for image filtering, Fuzzy Sets and Systems, vol. 161, no. 1, pp. 96-117, 2010. (ISSN: 0165-0114). ISI Web of Knowledge, IF 1.830 (2009 JCR Science Edition). http://dx.doi.org/.

[8] H.Hijazi, **O.Bazzi, A.Bigand** : " A new nonlinear discriminant analysis algorithm using a combined version of LDA and LLE", Congrès "ICPV2011 ", pp. 106 à 109, 18-23 juillet 2011, Las Vegas, USA.

[9] H.Hijazi, **O.Bazzi, A.Bigand** : " Out of samples extensions for SC-LLE, new non-linear dimensionality reduction algorithm", Congrès "ICCIT2013 ", 19-21 juin 2013, Beyrouth, Liban.

[10] **Joseph CONSTANTIN , Thèse soutenue le 11 mai 2001 à l'UPJV, professeur-associé invité au LISIC depuis 2012.**