HAL open science

Contributions à la classification et segmentation de séries temporelles par apprentissage statistique non supervisé ou guidé.

Émilie Poisson Caillault

To cite this version:

Émilie Poisson Caillault. Contributions à la classification et segmentation de séries temporelles par apprentissage statistique non supervisé ou guidé.. Intelligence artificielle [cs.AI]. Université du Littoral Côte d'Opale - ULCO, 2020. tel-03059280

HAL Id: tel-03059280

https://hal.science/tel-03059280v1

Submitted on 12 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

École doctorale n° 072 : Sciences Pour l'Ingénieur

Mémoire Présenté pour défendre le titre d' Habilitation à Diriger des Recherches

de l'Université du Littoral Côte d'Opale

Émilie Poisson Caillault

7 février 2020

Contributions à la classification et segmentation de séries temporelles par apprentissage statistique non supervisé ou guidé.

Garant : **André Bigand, HDR Université Littoral Côte d'Opale**

Jury

Jenny Benois-Pineau,	Professeur, Université de Bordeaux	Rapporteur
Sylvie Le Hégarat-Mascle,	Professeur, Université Paris-Sud	Rapporteur
Alain Rakotomamonjy,	Professeur, Université de Rouen	Rapporteur
Anne Renault,	Directrice Scientifique, IFREMER	Examinateur
Christian Viard-Gaudin,	Professeur, Université de Nantes	Examinateur
Alain Lefebvre,	HDR, Resp. IFREMER LER-BL	Invité

LISIC
ULCO
Calais

Table des matières

Table des matières iii

Liste des figures v

Liste des tableaux vii

1 Introduction 3
- 1.1 Série temporelle et événements caractéristiques . . . 5
- 1.2 Comparaison de séries ou de segments . . . 8
- 1.3 Segmentation non supervisée de séries temporelles . . . 12
- 1.4 Structuration du document . . . 13

2 Comparaison de séries temporelles 21
- 2.1 Métriques pour la comparaison de signaux . . . 22
- 2.2 Recherche de motifs pour l'imputation . . . 28
- 2.3 Classification et Partitionnement de séries . . . 41
- 2.4 Conclusions, Perspectives. . . . 48

3 Segmentation de séries temporelles 57
- 3.1 Introduction . . . 58
- 3.2 Segmentation guidée par les données . . . 62
- 3.3 Intégration des connaissances temporelles . . . 68
- 3.4 Segmentation d'événements ou zones isolés ou atypiques . . . 72
- 3.5 Conclusions, Perspectives. . . . 73

4 Conclusions 79

A Résumé des activités après thèse I
- A.1 CV simplifié . . . II
- A.2 Recherche . . . III
- A.3 Mobilité et Collaborations . . . V
- A.4 Encadrement doctoral et scientifique . . . V
- A.5 Projets, Contrats. . . . VIII
- A.6 Responsabilités collectives et Rayonnement scientifique . . . X
- A.7 Enseignement depuis 2006 . . . XIII

B Bibliographie Personnelle depuis 2006 XIX
- B.1 Publications dans des journaux JCR . . . XIX
- B.2 Chapitres d'ouvrages . . . XIX
- B.3 Publications dans des conférences internationales avec comité de sélection (full paper) . . . XIX
- B.4 Publications dans des conférences avec comité de sélection (sur résumé) et Communication depuis 2017 . . . XIX

B.5 Logiciels XIX

C Liste des acronymes **XXVII**

Liste des figures

1.1 Schéma de Dickey des processus phytoplanctoniques observables selon les échelles d'acquisition et les sources, tiré de **Karl2017dickey** 6
1.2 Série temporelle MAREL-Carnot de fluorescence en FFU (Fluoresceine FLuorescence Unit) de 2005 à nos jours, extraite du site web Coriolis Côtier. . . . 7
1.3 Superposition des années 2005 et 2006 des séries temporelles MAREL-Carnot de fluorescence. Le signal de 2005 est décalé de 40 FFU pour mieux visualiser les différences d'intensité et de forme des différentes efflorescences (eff.) printanières, estivales, automnales 7
1.4 Profil statistique hebdomadaire de la série temporelle MAREL-Carnot de fluorescence en FFU depuis 2005 ; Représentation par boîte de Tukey avec ajout de la moyenne représentée par les points rouge. 8
1.5 Comparaison de deux signaux q et r par Métrique point à point ou Métrique élastique 9

2.1 Signaux élémentaires artificiels et leurs statistiques de base. 24
2.2 Matrice de coût et chemin optimal d'appariement, tirée de (**GenTXWarper**) 25
2.3 Appariement et déformation DTW pour les signaux 'query' et 'ref1' 26
2.4 Illustration de la recherche d'un motif identique à la séquence Q précédant un T-Trou signifié ici par le symbole ?. 1- Construction de la séquence recherchée Q, 2-Comparaison des séquences proches par fenêtre glissante Qf, 3-Sélection de la fenêtre le plus similaire Qfs, 4- Imputation du T-Trou. 30
2.5 Prédictions sur 18 mois du taux d'humidité de Phu Lien selon différentes techniques comparées aux valeurs observées. 33
2.6 Illustration du processus d'imputation du signal de température de l'eau (WaterTemp) d'une sous-série issues des données Marel-Carnot. En rouge, la construction de la requête et la recherche d'une fenêtre similaire et en bleu le segment complété 34
2.7 Formulation de la fonction trapèze 34
2.8 Système expert calculant une similarité FBSM pondérée à partir de métriques de comparaison des signaux Q et R, et de règles floues (attention ici ED : Euclidean dissimilarity) 35
2.9 Partition floue des valeurs de similarités. 35
2.10 Complétion par recherche d'un phénomène passé et futur au T-Trou : 1- Construction des requêtes ; 2- Balayage des fenêtres similaires ; 3- Sélection de la ou les fenêtres admissibles ; 4 - fusion de ces fenêtres pour compléter le T-trou. 40
2.11 Comparaison des techniques de clustering sur plusieurs jeux de données générés tiré de Bob Baxley `http://rjbaxley.com`. (MB Kmeans : Minibatch Kmeans ; Aff. Prop. : Affinity Propagation, Agglo. Clust : Agglomerative Clustering, DBSCAN : Density Based Spatial Clustering of Applications with Noise, GMM : Gaussian Mixture Model) 42

2.12 Mesures de Rand Index sur différentes bases UCI (**Dua:2017**) et différents algorithmes de clustering non constraints (SC :spectral clustering) et contraints (SL basé kmeans contraint, cSC : constrained SC, FCSC Flexible Constrained SC, cPCA réduction par composantes principales contraintes, cLPP réduction contrainte par projection préservant les proximités locales entre points). . . . 46
2.13 Cytogramme 8D de 3 espèces phytoplanctoniques différentes. 47

3.1 Visualisation de la décomposition multiplicative tendance-cycle du signal de fluorescence mesurée par la station MAREL-Carnot (données Coriolis, IFREMER). 58
3.2 Visualisation des variabilités journalières des observations de fluorescence mesurée par la station MAREL-Carnot (données Coriolis, IFREMER). 59
3.3 Partitionnement k-means en deux groupes des données prétraitées MAREL-Carnot : Projection des clusters sur le signal de fluorescence mesurée par la station MAREL-Carnot de 2005 à 2010 et à gauche la répartition mensuelle toutes années confondues. 60
3.4 Segmentation par des approches de détection de changement de variance dans un jeu de données artificiel (à droite) et les données MAREL Carnot (à gauche). De haut en bas : Segmentation obtenue obtenue à partir des changements en moyenne et variance dans le signal y_3, Points de coupure par approche divisive multivariée, puis par approche agglomérative multivariée. 61
3.5 Processus de segmentation par classification dans l'espace spectral extrait des variables des observations. 63
3.6 Construction d'un hybride SC-HMM avec une extraction des états basées sur l'extraction des symboles. 66
3.7 Visualisation du clustering obtenu par SC-uHMM : dynamique des labels de 2005 à 2008, projection colorée des labels sur la fluorescence, répartition mensuelle des labels 67
3.8 Partitionnement des masses d'eau lors de legs de la campagne DYPHYMA : en haut la visualisation des trajets (LEG 1 à 3) ensuite les partitionnements par SC sur ces legs, chaque couleur désignant un label obtenu. 67
3.9 Exemple de valeur Q_i pour une fenêtre $T = 120$ sans ou avec données absentes. 68
3.10 Segmentation de la série formée des 4 paramètres de l'AOA signaux (en haut) du LEG 1 de la campagne DYPHYMA par clustering non contraint (image 2) et contrainte (3 et 4 en dessous) 69
3.11 M-SC : Clustering Spectral multi-niveau 72
3.12 Segmentation des données MAREL-Carnot sur la période 2005-2008 obtenu au niveau 3 de l'algorithme MSC). 73

4.1 Signal de salinité mesurée MAREL-Carnot avec de bas en haut les quantiles 10%, 25%, 50%, 75% et 90% représentés lignes pointillés et la moyenne en rouge. 82

Liste des tableaux

2.1 Métriques classiques de comparaison entre deux sous-séquences ou séquences univariées q et r. Les signes -/+/++ indiquent le niveau d'interprétabilité : (-) du non quantifié donc peu interprétable à (++) pleinement avec des bornes. . 23
2.2 Étude des métriques statistiques entre le signal query et l'ensemble des signaux élémentaires (en gras les valeurs respectant les bornes acceptables du critère) 24
2.3 Coûts d'appariement entre 'query' et signaux de références selon les métriques élastiques choisies (en gras les valeurs respectant les bornes acceptables du critère). 27
2.4 Métriques statistiques sur les signaux alignés après une recherche d'appariement avec le signal 'query', (en gras les valeurs pour lesquels les critères sont respectés) 27
2.5 Caractéristiques de jeux de données. (Tend. : tendance, Sais. : saisonnalité, CC-10 % : maximum de cross-corrélation entre le signal et des T-Trou de taille égale à 10 % du signal. 32
2.6 Indicateurs de performance des techniques d'imputation moyennés sur 10 à 50 T-Trous réalisés, T étant fixés à 10 % de la taille de chaque série (en gras, les meilleurs scores). 38
2.7 Indicateurs de qualité des techniques d'imputation sur des critères d'écart et de forme. Résultats moyennés sur un tirage de 5 T=2%-Trou de positions aléatoires. En gras, les résultats optimaux sont mis en évidence. 39
2.8 Temps de calcul moyen en seconde (s) pour compléter des T-trou dans la série "synthetic" selon la taille T et l'algorithme utilisé. 39

3.1 Indicateurs de performance de différents algorithmes de clustering pour segmenter la série artificielle. en gras sont repris les optimaux : ARI - Adjusted Rand Index (ARI), indices de Dunn et Silhouette (Sil.) et scores de précision (Acc.), événements détectés 73

A.1 Liste des enseignements de 2017 à 2019. XV
A.2 Liste des enseignements de 2014 à 2017). XVI
A.3 Liste des enseignements de sept. 2010 à août 2013 XVII
A.4 Liste des enseignements de jan. 2007 à août 2010 XVIII

Répondre ensemble à une problématique,
Faire émerger des thématiques,
Partager des valeurs,
Harmoniser nos heurs,
Discuter des idées,
Et rire de nos pensées.
Pour aboutir à ce travail d'équipe.

à Pierre et Célia,
à toutes les gemmes et leurs éclats,
précieux,
vaniteux,
de Bondar ou de Kalbur,
de Réalité ou d'Esprit,
de l'Océan,
surtout joyeux.

Emilie.

Chapitre 1

Introduction

« L'écriture,
chronique de nos actions,
état de nos réflexions,
ou début d'une nouvelle
aventure. »

EPC

Sommaire

1.1 Série temporelle et événements caractéristiques **5**
1.2 Comparaison de séries ou de segments **8**
1.2.1 Comparaison basée sur les instances du signal 9
1.2.2 Comparaison basée attribut 10
1.3 Segmentation non supervisée de séries temporelles **12**
1.3.1 Approche par morceaux 12
1.3.2 Approche par modèles 12
1.4 Structuration du document **13**

Contexte

Les travaux présentés dans ce manuscrit appartiennent aux domaines des **sciences des données** et de l'intelligence artificielle et plus particulièrement à l'**apprentissage statistique non supervisé ou faiblement guidé**. L'apprentissage non supervisé consiste à révéler une structuration dans un jeu de données et à les étiqueter sans connaissance du nombre d'étiquettes et/ou de la géométrie ou distribution des données appartenant à chaque étiquette. Nous appellerons indifféremment le mot étiquette ou label dans la suite de ce document. Le résultat de cet étiquetage automatique des données, c'est-à-dire sans expertise humaine spécifique, pourra servir de base d'apprentissage pour réaliser une classification dite supervisée d'un nouveau jeu de données (non déjà étiqueté). Les techniques d'apprentissage non supervisé - partitionnement, détection d'anomalies/intrus, réduction, *etc* - deviennent des éléments incontournables dans le contexte des mégadonnées (ère du big data) et d'aide à la décision et/ou modélisation.

Mes travaux de thèse s'inscrivaient dans ce seul cadre d'apprentissage supervisé et profond, avec la construction d'un système de reconnaissance en-ligne de l'écriture manuscrite cursive. Ce revirement entre les deux domaines, apprentissage supervisé à non supervisé, depuis ma prise de fonction de Maître de Conférences est brièvement expliqué avant d'introduire ses nouvelles orientations thématiques qui font l'objet de ce document.

Après l'obtention du diplôme d'Ingénieur en Systèmes Électroniques et Informatique Industrielle à Polytech'Nantes en 2001 et du Diplôme d'Études Approfondies (DEA équivalent

Master) d'Automatique Informatique Appliquée de l'Ecole Centrale de Nantes la même année, j'ai préparé ma thèse de doctorat de l'Université de Nantes (spécialité Automatique et Informatique Appliquée) soutenue en 2005. J'ai proposé plusieurs hybrides neuronaux (TDNN-SDNN) et un Système Hybride Neuro-Markovien pour la Reconnaissance de l'Écriture Manuscrite En-Ligne. Des critères d'apprentissage supervisés globaux (par approches conjointes discriminantes et maximum de vraisemblance) ont été proposés et validés sur plusieurs bases internationales étiquetées.

Avant mon arrivée effective en 2006, le laboratoire LASL - Laboratoire d'Analyse des Systèmes du Littoral - E.A. 2600 a subi une restructuration majeure. Par conséquent, j'ai intégré une équipe réduite à deux personnes, un professeur et un Maître de conférences dont la nouvelle orientation se concentrait sur la classification non supervisée d'un point de vue fondamental et sur le plan applicatif l' Environnement. En 2010, le LASL a fusionné pour devenir le laboratoire LISIC - Laboratoire d'Informatique Signal Image de Calais - EA 4491 divisé en quatre équipes dont IMAP - Image et Apprentissage dans laquelle les travaux ci-après s'inscrivent.

Ce revirement a donc été l'occasion d'enrichir ma vision des techniques dites d'apprentissage automatique (connu sous le nom de Machine Learning) et d'un point de vue applicatif parcourir un monde aquatique quelque peu invisible, insoupçonné et complexe à modéliser et rencontrer des chercheurs passionnés au travers de différents projets.

Résumé

Ce chapitre introductif présente plus précisément l'orientation et le cadre de mes recherches, à savoir expliquer des séries temporelles sans connaissance *a priori*, de la classification à la modélisation. Ces travaux sont axés et illustrés sur des applications réelles à la compréhension de phénomènes marins, notamment celle de la dynamique phytoplanctonique. Ils ont été le fruit d'une forte collaboration avec l'Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) depuis 2009 où j'ai effectué une délégation partielle entre 2015 et 2017 et des participations actives dans divers contrats (Interreg IVa 2 mers, H2020 JERICO-NEXT, ...).

À travers une revue illustrative et bibliographique du cadre de mes recherches, sont exposés les points suivants :

— une introduction générale à l'étude des séries temporelles, en particulier dans un contexte d'observation et de surveillance du milieu marin ;

— les grands cadres en classification et segmentation de séries ;

— le fil conducteur et le descriptif des chapitres suivants.

1.1 Série temporelle et événements caractéristiques

Les stations d'instrumentation ou bouées marines, telles en France le réseau MAREL - Mesures Automatisées en Réseau pour l'Environnement et le Littoral - déployé par IFREMER et le réseau COAST-HF - Coastal ocean observing system High frequency - géré par les infrastructures de recherche littorale et côtière (ILICO), offrent aujourd'hui un océan de données physico-chimiques à traiter. Formellement, elles fournissent une série d'observations $Y = (y_{ct})_{C \times T}$ de différents paramètres, appelées aussi variables y_c d'un capteur $\{c \in \mathbb{N}^* | c \leq C\}$ à différentes dates $\{t \in \mathbb{N}^* | t \leq T\}$. C et T sont des valeurs naturelles dénombrales. Chaque observation y_{ct} est un réel.

Emmanuel César et Bruno Richard dans leur manuel de cours [1] définissent ***une série temporelle*** en informatique comme

« une structure fondée sur les bases de données,
« fournissant ainsi le volume nécessaire d'information
« permettant de dresser une chronique historique des événements passés ».

Ces données collectées et mesurées généralement en routine sur de longues périodes T permettent ainsi une analyse statistique descriptive. Cette chronique, série chronologique Y_t, pourra ainsi être utilisée dans un but descriptif, explicatif ou de prévision de réalisations futures. Elle est généralement caractérisée comme la résultante d'une tendance Z_t représentant son évolution à long terme, des variations périodiques S_t (saisonnières et /ou cycliques) et des variations résiduelles aléatoires ϵ_t. Cette caractérisation suppose une certaine régularité et faible corrélation des trois composantes. Il conviendra de ne pas oublier que la donnée numérique existe seulement depuis un siècle et que notre évolution et celle de la dynamique phytoplanctonique ne se résume pas à ce siècle. Extraire de l'information dans une série temporelle, qu'elle soit mono ou multidimensionnelle (observations de plusieurs variables simultanément), dépend étroitement du contexte amont et visé : de la connaissance à la fois du processus d'acquisition et du processus que l'on souhaite identifier. À la caractérisation initiale, viennent se greffer des variations accidentelles d'intensité non négligeable contrairement à ϵ_t, que nous appellerons par la suite des **événements extrêmes** par leur amplitude ou leur rareté.

La dynamique phytoplanctonique présente ces caractéristiques de phénomènes extrêmes et/ou intermittents que nous illustrerons par la suite dans ce document. C'est un sujet d'attention fort et continu, puisque que le phytoplancton a été défini comme un des principaux indicateurs de la qualité des eaux et des écosystèmes dans la directive 2008/56/CE du Parlement européen et du Conseil du 17 juin 2008 appelée « Directive-cadre Stratégie pour le milieu marin » (DCSMM). En effet, le phytoplancton est d'une part le premier maillon de la chaîne alimentaire océanique constitué des cyanobactéries et microalgues présentes dans les eaux de surface et dérivant au gré des courants. Invisible à l'oeil nu, la prolifération de certains phytoplanctons est pourtant bien visible et parfois nuisible ou toxique. D'autre part, il est le premier producteur de l'oxygène terrestre et premier consommateur du dioxyde de carbone (plus de la moitié dans les deux cas), vulgarisé comme le « poumon de notre planète ». Pour se développer, le phytoplancton requiert l'énergie du soleil (lumière et température) et des éléments nutritifs comme le phosphore et le nitrate avec des conditions marines clémentes (faible turbidité). Il libérera alors de l'oxygène pendant sa photosynthèse.

Le suivi de la masse phytoplanctonique et de sa diversité est réalisé à partir d'échantillonnages spatiaux ou temporels de ces signaux dits Variables Océaniques Essentielles – ou Essential Ocean Variable (EOV) à la compréhension de la dynamique phytoplanctonique. Ces EOV surveillés sont des paramètres à la fois de forçage environnemental (luminosité, température, turbidité, concentration en nutriments, ...) et de réponse (concentration en

1. Module XML et Data Mining - Mars 2006, Les séries temporelles :georges.gardarin.free.fr/Surveys_DM/Survey_Time_Series.pdf

Oxygène dissous, quantité de pigments photosynthétiques pour distinguer les espèces, ...). Dans ce cadre, un grand nombre d'études d'analyse de tendance (M. Zhang et al. 2018) sont conduites à partir d'un seul signal : la fluorescence ou la Chlorophylle-a, considérées être un reflet de la biomasse phytoplanctonique. Quelques études temps-fréquences sont conduites à partir de combinaison de deux signaux/variables comme le couple Fluorescence/Température (Derot et al. 2015) ou Température/pourcentage de saturation en oxygène dissous (Huang et Schmitt 2014). Ces études sont basées sur des décompositions modales empiriques des signaux et leurs transformées de Hilbert dont les coefficients serviront à la comparaison. Peu de travaux s'intéressent au volume des données et à la richesse des événements qu'elles contiennent. Le mot volume réfère ici aussi bien à la fréquence d'acquisition qu'au contexte multicapteur/multiparamètre (avec $C \geq 2$).

La segmentation des séries temporelles en événements caractéristiques est un problème commun à plusieurs applications. Un événement peut être un élément essentiel ou non informatif. Dans une série (infra) journalière de chlorophylle-a ou fluorescence, les pics d'espèces dominantes sont souvent étroits et d'intensité variable. La période et forme de ces pics sont des événements informatifs, utiles pour comprendre la dynamique phytoplanctonique. Inversement, lors de la transcription d'un discours d'après un enregistrement, la claquement d'une porte sera considéré comme « bruit » c'est-à-dire inutile, non informatif.

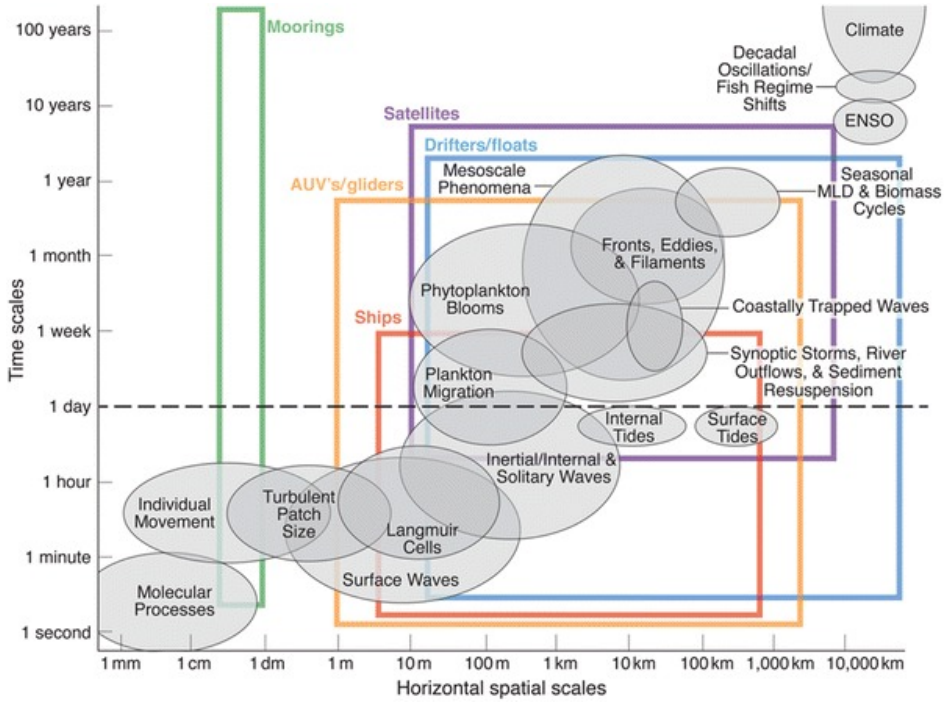


Figure 1.1 – Schéma de Dickey des processus phytoplanctoniques observables selon les échelles d'acquisition et les sources, tiré de Karl et Church 2017

Une adaptation du schéma de Dickey Figure 1.1 tiré de (Karl et Church 2017) renseigne sur les processus phytoplanctoniques observables (ellipses) selon les échelles et les types de sources (rectangles). Les résolutions d'acquisition temporelles et spatiales des observations mesurées permettent aujourd'hui d'appréhender le mécanisme d'initiation des efflorescences phytoplanctoniques (blooms). A partir d'observations satellites et de modèles numériques, une étude basée sur l'index de Shannon (Lévy et al. 2015) a montré que la biodiversité phytoplanctonique varie sur des échelles de temps de 1 à 30 jours et d'espace de 10 à 100 kilomètres.

La figure 1.2 illustre la fluorescence en unité FFU (Fluoresceine Fluorescence Unit) collectée toutes les vingt minutes depuis 2005 par la station MAREL-CARNOT en rade de Boulogne-sur-Mer, installée en avril 2004. Les variabilités des efflorescences, printanières,

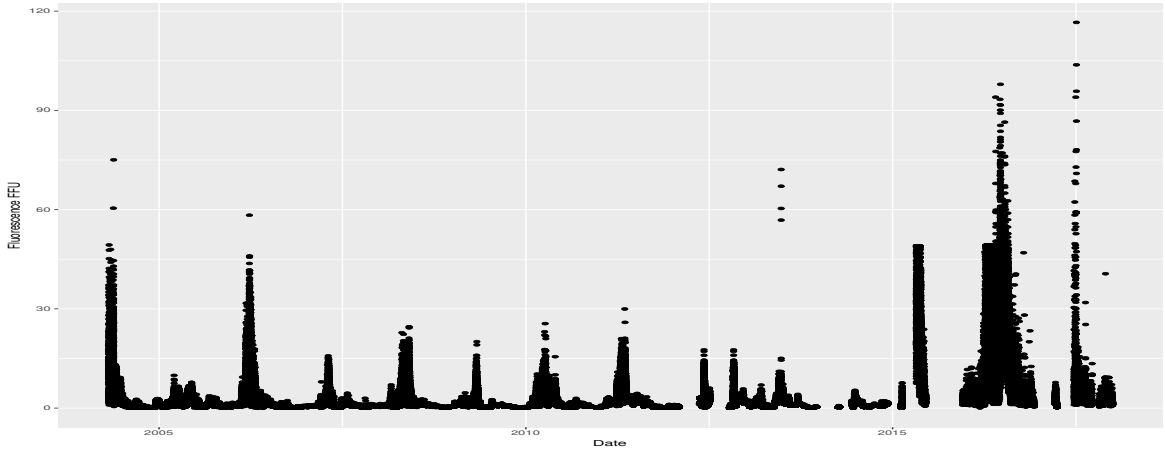


FIGURE 1.2 – Série temporelle MAREL-Carnot de fluorescence en FFU (Fluoresceine FLuorescence Unit) de 2005 à nos jours, extraite du site web Coriolis Côtier.

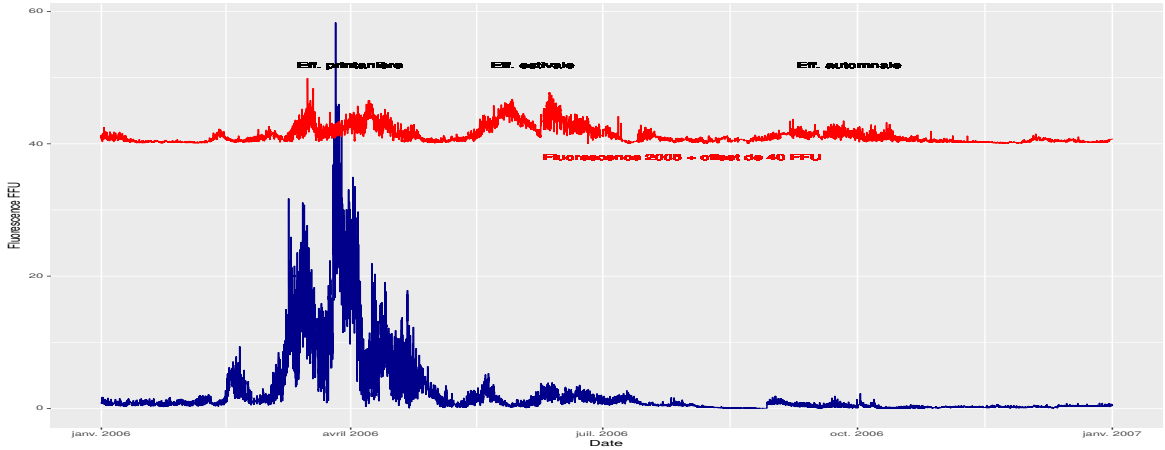


FIGURE 1.3 – Superposition des années 2005 et 2006 des séries temporelles MAREL-Carnot de fluorescence. Le signal de 2005 est décalé de 40 FFU pour mieux visualiser les différences d'intensité et de forme des différentes efflorescences (eff.) printanières, estivales, automnales

estivales ou automnales, sont mises en évidence par les différences d'intensité de 2005 à aujourd'hui, mais aussi de forme avec un zoom sur les années 2005 et 2006 figure 1.3. Elles sont aussi remarquables à travers le profil statistique hebdomadaire figure 1.4 de la fluorescence toutes années confondues. Il est à noter que les valeurs de FFU de 2012 à 2017 peuvent être limitées à la gamme maximale du capteur employé pour certaines périodes lors desquelles un plateau peut être observé (limite à 50 FFU). Le signal de fluorescence possède des fluctuations stochastiques très variables en intensité et en durée. Via des décompositions modales empiriques, il a été démontré que ces fluctuations ne sont pas du bruit et correspondent à des efflorescences elles-même de fluctuations variables en amplitude, durée et date de déclenchement.

Les données marines souffrent d'un manque d'information labellisée, les technologies actuelles ne permettent (pas encore) ni d'avoir la même granularité de traitement (infra-horaire à infra-journalier) ni d'avoir le niveau de labellisation souhaité pour maîtriser pleinement le processus du réseau phytoplanctonique à trophique (au mieux des grands groupes fonctionnels avec de large confusion et non espèces, THYSSEN et al. 2015). Pour interpréter ces séries, apporter des connaissances sur la dynamique phytoplanctonique et fournir des indicateurs d'alerte de qualité des eaux, il convient donc d'identifier les événements fréquents et/ou intermittents dans les données soit de segmenter la ou les séries collectées.

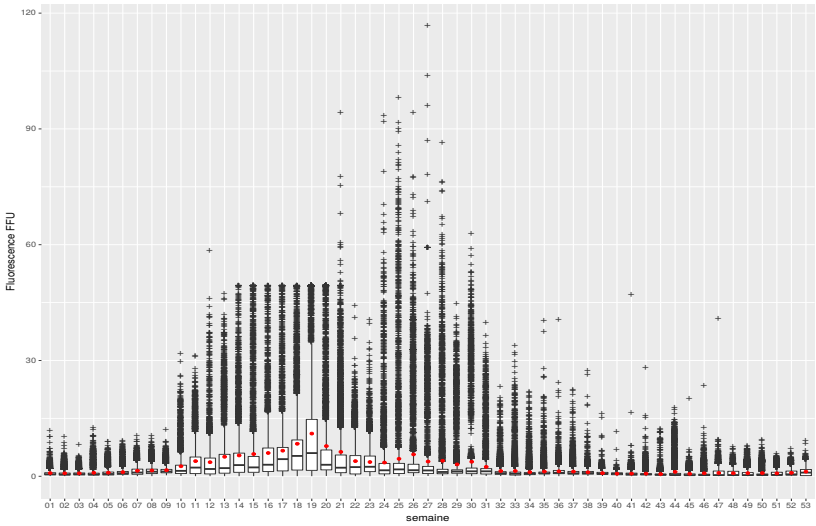


FIGURE 1.4 – Profil statistique hebdomadaire de la série temporelle MAREL-Carnot de fluorescence en FFU depuis 2005 ; Représentation par boîte de Tukey avec ajout de la moyenne représentée par les points rouge.

Le problème de segmentation d'une série temporelle peut alors être abordée sous différents paradigmes :

— « La vie est une succession de notes régulées. Notre musique n'est qu'un arrangement de quelques accords » . Il suffit alors d'**identifier des segments** temporels dans la série. Ces événements sont préalablement connus/appris. Cette labellisation de l'événement sera obtenu par **comparaison de séquences** à l'aide de métriques adéquates ou par **classification** supervisée.

— « Diem dies trudit, non similem sui - les années passent mais ne se ressemblent pas » . Elles sont composées d'événements rares, intermittents, extrêmes ou fréquents et parfois totalement destructurés. Cette hypothèse d'absence de connaissance *a priori* implique un processus de **segmentation non supervisée** : recherche des points de coupure, patron local, *etc.*

1.2 Comparaison de séries ou de segments

La comparaison de séries temporelles dépend étroitement d'une part de la représentation choisie et d'autre part de la métrique de (dis)similarité choisie. Deux grands courants se distinguent : une comparaison basée sur les instances de deux séries q et r ou une comparaison basée sur les attributs extraits de chaque série. Leurs schémas directeurs sont les suivants :

Soient deux vecteurs r et q à valeurs dans $\mathbb{R}$ de taille respective T_r et T_q

Comparaison des instances :
1- $(q, r) \rightarrow$ Appariement linéaire ou élastique $\rightarrow (qa, ra)$; qa et ra de taille $T_a \geq \max(T_q, T_r)$ 2- Calcul d'une métrique : $d(qa, ra)$
Comparaison des attributs :
1- signal $r \rightarrow$ Représentation Attributs $\rightarrow f(r) \in \mathbb{R}^D$ Sélection /réduction $\rightarrow f'(r) \in \mathbb{R}^d, d \leq D$ 2- signal $q \rightarrow$ Représentation Attributs $\rightarrow f(q) \in \mathbb{R}^D$ Sélection/réduction $\rightarrow f'(q) \in \mathbb{R}^d, d \leq D$ 3- Calcul d'une métrique : $d(f'(r), f'(q))$

1.2.1 Comparaison basée sur les instances du signal

Deux séries univariées q et r seront considérées proches si la distance/dissimilarité $d(q, r)$ est faible. Pour aider à l'interprétation de cette dissimilarité, une normalisation des séries (normalisation min-max ou z-score) est parfois opérée, réduisant les signaux à une amplitude entre $[0; 1]$. Les paramètres de cette normalisation devant être connus ou pertinents, telles les valeurs minimum et maximum, ou la moyenne et écart-type.

Distance point à point

Deux séries univariées q et r seront considérées proches si la distance/dissimilarité entre leurs couples de valeurs à l'instant t, $(q[t], r[t])$ est faible. Selon l'application, différents indicateurs d'écart sont préférés : l'erreur quadratique moyenne normalisée (NMSE - normalized mean square error) ou sa racine carré (RMSE - root MSE), l'erreur absolue moyenne (MAE) ou des critères d'éloignement par rapport à une enveloppe acceptable via les biais fractionnels ou géométriques. Quelle que soit la normalisation du signal choisi, ces métriques supposent que les deux séries sont parfaitement alignées selon l'axe temporel et de même longueur $T_r = T_q$.

Métrique élastique

Deux séries univariées q et r seront considérées proches si la dissimilarité entre les couples des valeurs de leurs signaux alignés à l'instant t est faible. Soient q_a et r_a les signaux alignés, alors $d(q_a[t], r_a[t])$ est faible.

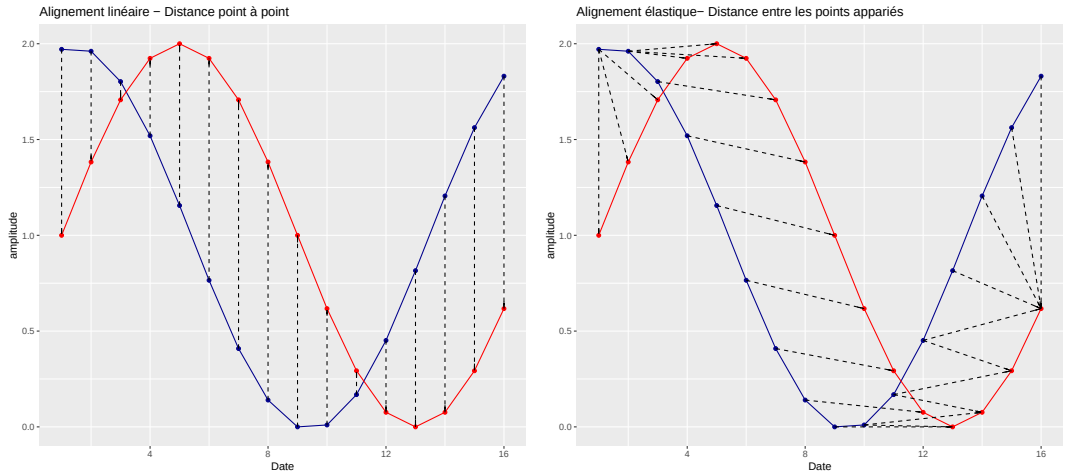


FIGURE 1.5 – Comparaison de deux signaux q et r par Métrique point à point ou Métrique élastique

La figure 1.5 permet de comprendre la différence entre une métrique point à point (alignement linéaire) et une métrique de signaux appariés avec une tolérance de déformation temporelle. Le concept d'alignement/appariement élastique de signaux a été introduit par Sakoe et Chiba en 1978 pour la reconnaissance vocale de mots (SAKOE et CHIBA 1978). Les syllabes d'un mot peuvent être prononcées plus lentement et avec une intonation différente, il convient donc de construire une métrique tolérante à ces déformations. Un critère de coût de déformation (en anglais DTW pour Dynamic Time Warping) est optimisé pour obtenir un alignement global optimal entre les séquences. Ainsi chaque élément de chaque séquence est associé à au moins un élément de l'autre séquence en minimisant les coûts d'association sans retour en arrière possible. Ce coût d'association d appartient généralement à la famille des Lp-normes. L'algorithme initial DTW, transcrit à l'algorithme 1, est basé sur la construction d'une matrice de coût DTW remplie de manière ordonnée temporellement. DTW[i,j] correspond au coût intégré entre le point à l'instant i de q et le point j de r, soit la somme

de la distance $d(q[i], r[j])$ avec le minimum des coûts de déplacements précédents possibles (horizontal,vertical ou diagonal). Le coût d'appariement des points finaux des deux séries $DTW[T_r, T_q]$ est la distance nécessaire pour apparier les deux signaux avec une tolérance de déformation temporelle de taille w.

Algorithme 1 Algorithme DTW - Sakoe et Chiba

Require: int $w \geq 1$, r : array $[1..T_r]$, q : array $[1..T_q]$
  DTW : array $[0..T_r, 0..T_q] \leftarrow 0$
  for i=1..T_r **do**
    DTW[i, 0] $\leftarrow$ infinity
  end for
  for i=1..T_q **do**
    DTW[0, i] $\leftarrow$ infinity
  end for
  for i=1..T_r **do**
    for j= $\max(1, i-w)..\min(T_q, i+w)$ **do**
      cost $\leftarrow$ d(r[i],q[j])
      DTW[i, j] $\leftarrow$ cost + min(DTW[i-1,j],DTW[i,j-1],DTW[i-1,j-1])
    end for
  end for
  return DTW[T_r,T_q]

Cet algorithme a été déployé et éprouvé dans de nombreux domaines. Plusieurs versions d'optimisation calculatoire ont été proposées dont la principale est celle de l'estimation préalable d'une enveloppe acceptable (E. Keogh et Ratanamahatana 2005a ; Dau et al. 2018).

Différentes variantes ont été proposées afin d'améliorer l'alignement des séries, en respectant plus finement les formes locales du signal : Derivative DTW (DDTW) modifiant la série en utilisant sa dérivée première (E. J. Keogh et Pazzani 2001) ; Adaptive Feature Based DTW (AFBDTW) (Xie et Wiltgen 2010) intégrant les dérivées locales, des attributs locaux et globaux en chaque point dans le calcul de la distance d. Chen et al. Y. Chen, Hu et al. 2013 ont proposé une nouvelle mesure de similarité DTW-D en normalisant le coût global DTW par la distance euclidienne entre les deux séries afin de mieux discriminer les signaux de faibles amplitudes. Nous avons proposé une comparaison de ces différentes variantes dans un contexte d'imputation de séries climatiques et marines (T.T.H. Phan et al. 2017).

Le paragraphe ci-dessus ne présente qu'une partie des métriques élastiques entre deux séries (pour une liste plus exhaustive se référer à Mitsa 2010 ; Wang et al. 2013). Nous retiendrons en particulier celle basée sur une sélection/réduction préalable de leur plus longue séquence commune (LCSS - Longest Common SubSequence Lin et Shim 1995 Górecki 2018).

D'autres auteurs ont cherché à conserver une métrique avec les propriétés d'une distance telles TWED (Time Warp Edit Distance) Marteau 2007 variante de DTW conservant la propriété de l'inégalité triangulaire, ERP : Edit distance with Real Penalty L. Chen et Ng 2004 ou une combinaison de ces métriques élastiques Lines et Bagnall 2015.

1.2.2 Comparaison basée attribut

Les approches basées sur l'extraction d'attributs issus d'une représentation ont largement été prônées pour la comparaison de deux séries de grandes tailles (T grand) ou la compression et indexation de larges bases (Fu 2011). Aujourd'hui ce choix est surtout lié à la finalité, car les outils précédents ont été étendus, optimisés pour de larges séquences.

Représentation par transformée

Les représentations par Transformée de Fourier Discrète (DFT : Discrete Fourier Transform), Transformée par ondelettes discretes (DWT : Discrete Wavelet Transform) ou Transformée de Hilbert–Huang (HHT) sont largement utilisées pour étudier les similitudes entre deux séries (climatologie : Z. Zhang et Moore 2015) ou pour réaliser une tâche de classification d'une ensemble de séries mais avec des tailles de séries courtes (classification de cytogrammes phytoplanctoniques par FFT : Malkassian et al. 2011) . La décomposition en valeurs singulières (SVD- Singular value decomposition) a aussi été explorée pour compresser la taille des séries dans un but de stockage, d'indexation ou de comparaison de l'évolution de deux séries (climatologie : Chl-a/temperature de l'eau en surface Kang et al. 2017). A partir de ces transformées, une grande diversité d'attributs peut être calculée telles les tendances, saisonnalités, les caractéristiques de forme comme le coefficient d'aplatissement (kurtosis) ou coefficient d'asymétrie (skewness). Le choix des coefficients, choix du nombre de modes ou d'ondelettes à utiliser ou des attributs pertinents sont autant de paramètres à considérer pour améliorer la comparaison et la classification entre séries.

Autres Représentations

Les représentations par morceaux se sont aussi largement développées dans un but principal de compression et d'accélération du calcul de similarité ou de la classification entre séries. Les séries sont alors découpées en segments ; à chaque segment est assigné une valeur ou un codage. Dans la PAA (Piecewise Aggregate Approximation Guo, Li et Pan 2010), les segments sont de taille fixe et représentés par la valeur moyenne de ses valeurs ou un ensemble de statistiques. Keogh et al. ont proposé une version combinée de la PAA avec une similarité DTW E. Keogh et Ratanamahatana 2005b. Les divers codages des segments (SDL, CAPSUL, iSAX) sont détaillés (Mitsa 2010) tel l'ensemble des symboles (« Up, up, stable, zero, down, DOWN »). Très intéressants dans un contexte de réduction de la dimensionnalité, ils ne permettent pas ici d'interpréter la dynamique des efflorescences phytoplanctoniques illustrée à la figure 1.2. En effet, l'intensité des segments n'est pas différenciée or l'abondance phytoplanctonique a un rôle informatif prépondérant sur la composition, diversité et richesse des espèces à l'instant t et celles consécutives. Les modèles de Markov cachés (Hidden Markov Model (HMM) Rabiner 1989) engendrent une représentation des séries temporelles comme des graphes d'événements appelés, dans le jargon markovien, des états dont les paramètres sont inconnus. A partir d'une base d'apprentissage, les paramètres dynamiques (probabilité d'émission d'un état, probabilité de transition entre états, probabilité de commencer par tel état) peuvent être déterminés par l'algorithme de Baum-Welch ou Viterbi via une estimation itérative par maximum de vraisemblance. Le nombre et la caractérisation des états sont des paramètres à déterminer. Cette modélisation sera préférée pour interpréter la dynamique d'une série, et non dans un but de comparaison ou de classification.

Combinaison et Sélection

Fulcher et Jones 2014 ont extrait une dizaine de milliers d'attributs combinant des attributs issus de ces transformées et de statistiques liées à la distribution des séries, corrélations linéaires, mesures d'entropies, . . . , puis ils ont sélectionnés les plus discriminants par apprentissage sur une base de séries étiquetées. Leur étude a porté sur une vingtaine de bases issues de la banque de données UCR Times Séries Classification (Y. Chen, E. Keogh et al. 2015) avec une comparaison avec un classifieur 1-NN DTW (figure 5 et table 1 de l'article cité notamment). Les attributs discriminants dépendent de la base utilisée. Leur étude ne permet pas de statuer sur le choix de ce type d'approche vis-à-vis d'une approche 1-NN DTW. Cependant elle permet d'offrir aux parties intéressées les propriétés dominantes des séries. Le classifieur élementaire par plus proche voisin (1-ppv ou 1-Nearest Neighbor) a montré ses

preuves dans un grand nombre de domaines, notamment avec une combinaison DTW (Y. Chen, E. Keogh et al. 2015 ; Rakthanmanon et al. 2012 ; Ding et al. 2008).

Une fois la distance choisie, l'indexation de séquences/séries revient ainsi à rechercher toutes les séquences y_q proches de la série de référence y_r telle que $d(y_q, y_r) \leq \epsilon$ si celle-ci est connue, ou à identifier des ensembles de séquences proches.

Nous reviendrons sur ces deux grands volets de comparaison avec un intérêt particulier pour les métriques élastiques très adaptés aux signaux à fortes variabilités de déclenchement comme les efflorescences phytoplanctoniques.

1.3 Segmentation non supervisée de séries temporelles

Identifier des événements dans une série requiert une étape de segmentation. Cette notion d'événements et de la technique à utiliser dépend étroitement de la finalité visée. Un événement phytoplanctonique peut être à la fois un segment non redondant, comme la production d'une algue toxique jamais apparue précédemment et un segment intermittent, revenant régulièrement mais pas forcément à la même période ni avec la même biomasse. Sa durée, sa forme sont autant de paramètres fort variables et parfois inconnus.

Identifier des événements dans une série sans connaissance *a priori* ni de leur forme ni de leur phénologie revient à segmenter celle-ci de manière aveugle. Nous présentons les deux grandes classes de segmentation basées sur un découpage et étude par morceau (fenêtre) ou sur une identification de patrons locaux appelés aussi modèles.

1.3.1 Approche par morceaux

La représentation d'une série temporelle par morceaux, dans la littérature PLR - Piecewise Linear Representation a été introduite initialement par Shatkay et Zdonik en 1996 Shatkay et Zdonik 1996 et largement déclinée. Le principe est de diviser la série en multiples segments soit

— par un découpage selon des points de coupure : points d'importance (Pratt et Fink 2002) ; points d'intérêt (PIP -Perception Interest Points Tsinaslanidis et Kugiumtzis 2014) ; points de ruptures (EVT - Extreme values Theory Kuswanto, Andari et Oktania Permatasari 2015) ; ...

— soit par une étude de fenêtres. Ces études par fenêtres sont obtenues soit par une segmentation de taille fixe (Van Hoan, Huy et L.C. 2017), division successive de la série ou à l'inverse accroissement de la taille de la fenêtre reprises (méthodes reprises dans Last, Kandel et Bunke 2004) ou encore fenêtres mobiles (Längkvist, Karlsson et Loutfi 2014).

Quelle que soit la technique de découpage utilisée, la bibliothèque de segments obtenue est ensuite partitionnée en K groupes pour isoler des anomalies ou des comportements caractéristiques. Un tutoriel de Muen and Keogh (Mueen et E. Keogh 2017) illustre le partitionnement d'une série en classe de segments caractéristiques par création d'une matrice de profils.

Dans le contexte de bases de signaux échantillonnés toutes les 20 minutes avec des phénomènes phytoplanctoniques parfois infra-journaliers depuis les années 2000, cette approche de segmentation engendre un coût calculatoire exponentiel et un effort d'étiquetage expert trop importante malgré l'étape de réduction en groupes. Cette conclusion sera reprise au chapitre 3 de ce manuscrit.

1.3.2 Approche par modèles

Par ailleurs, l'expert en eutrophisation n'a pas l'habitude d'analyser visuellement une série par ce type de découpage. Il va plutôt identifier des formes courantes (palier, cloches,

pics ou dans le jargon associé période non productive, début d'efflorescences, maintien, fin, efflorescence automnale ...).

Une série peut alors être étudiée comme la somme de modèles ou lois dont les paramètres sont obtenus par algorithme d'Espérance-Maximisation (EM, expectation-maximization) proposé par (Dempster, Laird et Rubin 1977). Dans (Emonet, Varadarajan et Odobez 2014), une série est modélisée à partir de processus de Dirichlet. Dans (Poisson Caillault et Lefebvre 2017), nous avons proposé de décomposer une série comme une succession de courbes gaussiennes avec un faible recouvrement temporel. Une série peut aussi être représentée par un graphe d'événements avec une modélisation HMM (K. Rousseeuw et al. 2015 ; Dias, Vermunt et Ramos 2015) apportant des informations dynamiques supplémentaires.

1.4 Structuration du document

Le recours aux graphes, aussi bien dans la représentation d'une série temporelle que dans la classification de données, a été une motivation forte dans l'ensemble de mes travaux présentés ici. La représentation visuelle d'une information et de la finalité d'un algorithme est (1) une aide précieuse pour comprendre le mécanisme associé et (2) un outil d'interprétation riche permettant de rapprocher plus aisément des communautés scientifiques de langages différents (aparté amusé sur des termes, oppositions si souvent entendus : informaticien/traiteur de signaux, informaticien/biologiste).

La suite du manuscrit expose une partie des travaux de recherche que j'ai pu réaliser au cours d'encadrement de thèses, stages de master ou en collaboration avec d'autres collègues dans divers projets locaux ou internationaux. La démarche générale adoptée est guidée par la conjecture suivante : *« Segmenter une série nécessite une parfaite connaissance de celle-ci et, des outils d'extraction pertinents. »* qui se traduit par trois conditions suivantes :

1. Les informations aberrantes ou absentes dans la série devront être identifiées et traitées pour avoir une compréhension entière du phénomène.
2. Les segments obtenus appelés aussi événements, devront avoir un sens pour l'application.
3. Les segments devront être catégorisés.

Nous supposerons que la série à étudier possède des segments intermittents que nous souhaitons être catégorisés ensemble et des événements extrêmes ou rares dans des catégories isolées. De cette problématique et hypothèse, l'approche par classification et comparaison de segments est naturelle.

Le manuscrit est donc divisé en deux chapitres principaux l'un traitant de la comparaison de segments et l'autre du thème général de la segmentation et modélisation de séries en segments par classification non supervisée avec un intérêt particulier pour les approches spectrales et les métriques élastiques.

Le chapitre 2 expose plusieurs de nos contributions sur la comparaison de motifs (sous-séquences) pour différentes applications. La première section 2.1 introductive donne un aperçu des mesures utilisées dans la comparaison de signaux et de leur pertinence. La section 2.2 est dédiée à la recherche de motifs pour compléter des sous-séquences absentes dans une série. En premier lieu, un point de vue sur les techniques d'imputation est apporté. Elle inscrit nos travaux dans les sciences de la fouille de données et de l'incertain. Notre démarche est basée sur l'exploitation de la redondance existante d'un phénomène et les métriques élastiques pour combler de larges trous dans les séries où les techniques d'interpolation échouent à conserver la forme des événements la constituant. Trois algorithmes sont proposés dans cette tâche. Nous montrons que ces algorithmes peuvent également être employés dans des tâches de prédiction court et moyen terme d'une série. La section 2.3 est dédiée aux partitionnements de segments (séries) en vue d'aider un expert dans une tâche de labellisation et limiter la

mise en oeuvre de celle-ci qui selon le type d'applications peut être fastidieuse, onéreuse ou hors-sujet vis-vis de la quantité d'information à traiter. Un point de vue personnel est apporté sur les techniques de clustering et précisément les motivations à creuser les approches spectrales. Dans la section 2.3.2, nous repartons du formalisme général de clustering spectral pour intégrer des connaissances par paire d'objets. Ces connaissances sont plus aisées à obtenir d'un expert, voire d'un non spécialiste de l'application. La section 2.4 conclut en résumant les principaux résultats et en proposant des perspectives à ces travaux.

Le chapitre 3 développe nos recherches sur la segmentation d'une série mono- ou multivariée en événements caractéristiques sans connaissance ou peu sur le phénomène et sa décomposition visées. La section introductive 3.1.1 part d'une application phare, la caractérisation des états environnementaux d'une masse d'eau pour illustrer les sources de difficultés potentielles des approches de segmentation par détection de rupture. Ce bilan montre l'intérêt d'utiliser et de proposer une approche totalement guidée par les données et non par la dimension spatio-temporelle dans la section 3.2. Nous avons proposé une modélisation par graphe : une série peut être vue comme une suite d'événements (noeuds du graphe) avec une dynamique particulière de passage entre ces événements (arcs). Nous nous sommes appuyés sur un formalisme ergodique via un modèle de Markov Caché d'ordre 1 dont les états sont déterminés automatiquement par classification spectrale, section 3.2.1. L'apprentissage totalement non supervisé et direct par opposition aux techniques usuelles d'Expectation-Maximisation et de programmation dynamique est détaillé en section 3.2.2. Nous étendons ensuite ce modèle à une segmentation guidée à la fois par les données et la connaissance du phénomène observé (vitesse de changement, phénologie) en reprenant le même formalisme de classification spectrale contrainte vu au chapitre 2.2. Un dernier volet, à la section 3.4, est consacré à la détection d'événements dit extrêmes par leur forme ou leur rareté d'apparition. Nous proposons une nouvelle architecture de classification spectrale divisive permettant d'obtenir une segmentation hiérarchisée de nos événements d'un point de vue global à une approche fine en événement isolé, extrême. La dernière section du chapitre 3 établit une conclusion qui résume les principaux résultats de celui-ci et propose quelques perspectives.

Le manuscrit se termine par un court chapitre 4 qui reprend des travaux issus de quatre thèses encadrées et expose les prochains axes de travaux en continuité et nouveaux au-delà des perspectives déjà exprimées dans chaque chapitre.

Bibliographie

L. Chen et Ng 2004

Chen, Lei et Raymond Ng (2004). « On the marriage of lp-norms and edit distance ». In : *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, p. 792–803.

Y. Chen, Hu et al. 2013

Chen, Yanping, Bing Hu et al. (2013). « DTW-D: Time Series Semi-supervised Learning from a Single Example ». In : *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA : ACM, p. 383–391. isbn : 978-1-4503-2174-7. doi : 10.1145/2487575.2487633. url : http://doi.acm.org/10.1145/2487575.2487633.

Y. Chen, E. Keogh et al. 2015

Chen, Yanping, Eamonn Keogh et al. (2015). *The UCR Time Series Classification Archive*. www.cs.ucr.edu/~eamonn/time_series_data/.

Dau et al. 2018

Dau, Hoang Anh et al. (2018). « Optimizing dynamic time warping's window width for time series data mining applications ». In : *Data mining and knowledge discovery* 32.4, p. 1074–1120.

Dempster, Laird et Rubin 1977

Dempster, A. P., N. M. Laird et D. B. Rubin (1977). « Maximum likelihood from incomplete data via the EM algorithm ». In : *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1, p. 1–38. url : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.4884.

Derot et al. 2015

Derot, J. et al. (2015). « Long-term high frequency phytoplankton dynamics, recorded from a coastal water autonomous measurement system in the eastern English Channel ». In : *Continental Shelf Research* 109, p. 210–221. issn : 0278-4343. doi : https://doi.org/10.1016/j.csr.2015.09.015. url : http://www.sciencedirect.com/science/article/pii/S0278434315300674.

Dias, Vermunt et Ramos 2015

Dias, J.G., J.K. Vermunt et S. Ramos (2015). « Clustering financial time series: New insights from an extended hidden Markov model ». In : *European Journal of Operational Research* 243.3, p. 852–864. issn : 0377-2217. doi : https://doi.org/10.1016/j.ejor.2014.12.041. url : http://www.sciencedirect.com/science/article/pii/S0377221714010595.

Ding et al. 2008
Ding, Hui et al. (2008). « Querying and mining of time series data: experimental comparison of representations and distance measures ». In : *Proceedings of the VLDB Endowment* 1.2, p. 1542–1552.

Emonet, Varadarajan et Odobez 2014
Emonet, R., J. Varadarajan et J-M. Odobez (2014). « Temporal Analysis of Motif Mixtures Using Dirichlet Processes ». In : *IEEE Trans. Pattern Anal. Mach. Intell.* 36.1, p. 140–156. doi : 10.1109/TPAMI.2013.100. url : https://doi.org/10.1109/TPAMI.2013.100.

Fu 2011
Fu, Tak-chung (2011). « A review on time series data mining ». In : *Engineering Applications of Artificial Intelligence* 24.1, p. 164–181.

Fulcher et Jones 2014
Fulcher, B. D. et N. S. Jones (2014). « Highly Comparative Feature-Based Time-Series Classification ». In : *IEEE Transactions on Knowledge and Data Engineering* 26.12, p. 3026–3037. issn : 1041-4347. doi : 10.1109/TKDE.2014.2316504.

Górecki 2018
Górecki, Tomasz (2018). « Classification of time series using combination of DTW and LCSS dissimilarity measures ». In : *Communications in Statistics - Simulation and Computation* 47.1, p. 263–276. doi : 10.1080/03610918.2017.1280829.

Guo, Li et Pan 2010
Guo, Chonghui, Hailin Li et Donghua Pan (2010). « An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining ». In : *Knowledge Science, Engineering and Management*. Sous la dir. d'Yaxin Bi et Mary-Anne Williams. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 234–244. isbn : 978-3-642-15280-1.

Huang et Schmitt 2014
Huang, Y. et F.G. Schmitt (2014). « Time dependent intrinsic correlation analysis of temperature and dissolved oxygen time series using empirical mode decomposition ». In : *Journal of Marine Systems* 130, p. 90–100. doi : 10.1016/j.jmarsys.2013.06.007.

K. Rousseeuw et al. 2015
K. Rousseeuw et al. (2015). « Hybrid Hidden Markov Model for Marine Environment Monitoring ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.1. Impact Factor:2.15, p. 204–213. issn : 1939-1404. doi : 10.1109/JSTARS.2014.2341219.

Kang et al. 2017
Kang, Xianbiao et al. (2017). « An improved ENSO simulation by representing chlorophyll-induced climate feedback in the NCAR Community Earth System Model. » In : *Scientific Reports* 7, p. 17123.

Karl et Church 2017
Karl, David M. et Matthew J. Church (2017). « Ecosystem Structure and Dynamics in the North Pacific Subtropical Gyre: New Views of an Old Ocean ». In : *Ecosystems* 20.3, p. 433–457. issn : 1435-0629. doi : 10.1007/s10021-017-0117-0. url : https://doi.org/10.1007/s10021-017-0117-0.

E. J. Keogh et Pazzani 2001
Keogh, Eamonn J. et Michael J. Pazzani (2001). « Derivative Dynamic Time Warping ». In : *In First SIAM International Conference on Data Mining (SDM'2001)*.

E. Keogh et Ratanamahatana 2005a
Keogh, Eamonn et Chotirat Ann Ratanamahatana (2005a). « Exact indexing of dynamic time warping ». In : *Knowledge and Information Systems* 7.3, p. 358–386. ISSN : 0219-3116. DOI : 10.1007/s10115-004-0154-9. URL : https://doi.org/10.1007/s10115-004-0154-9.

E. Keogh et Ratanamahatana 2005b
— (2005b). « Exact indexing of dynamic time warping ». In : *Knowledge and information systems* 7.3, p. 358–386.

Kuswanto, Andari et Oktania Permatasari 2015
Kuswanto, H., S. Andari et E. Oktania Permatasari (2015). « Identification of Extreme Events in Climate Data from Multiple Sites ». In : *Procedia Engineering* 125. Civil Engineering Innovation for a Sustainable, p. 304–310. ISSN : 1877-7058. DOI : 10.1016/j.proeng.2015.11.067.

Längkvist, Karlsson et Loutfi 2014
Längkvist, M., L. Karlsson et A. Loutfi (2014). « A review of unsupervised feature learning and deep learning for time-series modeling ». In : *Pattern Recognition Letters* 42, p. 11–24. ISSN : 0167-8655. DOI : https://doi.org/10.1016/j.patrec.2014.01.008.

Last, Kandel et Bunke 2004
Last, Mark, Abraham Kandel et Horst Bunke (2004). *Data Mining in Time Series Databases*. WORLD SCIENTIFIC. DOI : 10.1142/5210.

Lévy et al. 2015
Lévy, Marina et al. (2015). « The dynamical landscape of marine phytoplankton diversity ». In : *Journal of The Royal Society Interface* 12.111. ISSN : 1742-5689. DOI : 10.1098/rsif.2015.0481. URL : http://rsif.royalsocietypublishing.org/content/12/111/20150481.

Lin et Shim 1995
Lin, Rake& Agrawal King-lp et Harpreet S Sawhney Kyuseok Shim (1995). « Fast similarity search in th presence of noise, scaling and translation in time-series databases ». In : *Proc. of the 21st VLDB conference, Zurich, Switzerland, 1995*. URL : https://ci.nii.ac.jp/naid/80008714321/en/.

Lines et Bagnall 2015
Lines, Jason et Anthony Bagnall (2015). « Time Series Classification with Ensembles of Elastic Distance Measures ». In : *Data Min. Knowl. Discov.* 29.3, p. 565–592. ISSN : 1384-5810. DOI : 10.1007/s10618-014-0361-2. URL : http://dx.doi.org/10.1007/s10618-014-0361-2.

Malkassian et al. 2011
Malkassian, Anthony et al. (2011). « Functional analysis and classification of phytoplankton based on data from an automated flow cytometer ». In : *Cytometry Part A* 79A.4, p. 263–275. DOI : 10.1002/cyto.a.21035. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.21035. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.21035.

Marteau 2007
Marteau, Pierre-Francois (2007). « Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching ». In : *CoRR* abs/cs/0703033. arXiv : cs/0703033. URL : http://arxiv.org/abs/cs/0703033.

Mitsa 2010 Mitsa, Theophano (2010). *Temporal Data Mining*. 1st. Chapman et Hall/CRC.

Mueen et E. Keogh 2017
Mueen, Abdullah et Eamonn Keogh (2017). *Time Series data Mining Using the Matrix Profile: A Unifying View of Motif Discovery, Anomaly Detection, Segmentation, Classification, Clustering and Similarity Joins*. tutorial of KDD2017.

Poisson Caillault et Lefebvre 2017
Poisson Caillault, É. et A. Lefebvre (2017). « Towards Chl-a bloom understanding by EM-based unsupervised event detection ». In : *OCEANS 2017 - Aberdeen*, p. 1–5. DOI : 10.1109/OCEANSE.2017.8084597.

Pratt et Fink 2002
Pratt, Kevin B. et Eugène Fink (2002). « Search For Patterns in Compressed Time Series ». In : *International Journal of Image and Graphics* 02.01, p. 89–106. DOI : 10.1142/S0219467802000482.

Rabiner 1989
Rabiner, Lawrence R (1989). « A tutorial on hidden Markov models and selected applications in speech recognition ». In : *Proceedings of the IEEE* 77.2, p. 257–286.

Rakthanmanon et al. 2012
Rakthanmanon, Thanawin et al. (2012). « Searching and Mining Trillions of Time Series Subsequences Under Dynamic Time Warping ». In : *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. Beijing, China : ACM, p. 262–270. ISBN : 978-1-4503-1462-6. DOI : 10.1145/2339530.2339576. URL : http://doi.acm.org/10.1145/2339530.2339576.

Sakoe et Chiba 1978
Sakoe, H. et S. Chiba (1978). « Dynamic programming algorithm optimization for spoken word recognition ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1, p. 43–49. ISSN : 0096-3518. DOI : 10.1109/TASSP.1978.1163055.

Shatkay et Zdonik 1996
Shatkay, H. et S. B. Zdonik (1996). « Approximate queries and representations for large data sequences ». In : *Proceedings of the Twelfth International Conference on Data Engineering*, p. 536–545. DOI : 10.1109/ICDE.1996.492204.

Thyssen et al. 2015
Thyssen, M. et al. (2015). « High-resolution analysis of a North Sea phytoplankton community structure based on in situ flow cytometry observations and potential implication for remote sensing ». In : *Biogeosciences* 12.13, p. 4051–4066. DOI : 10.5194/bg-12-4051-2015. URL : https://hal-amu.archives-ouvertes.fr/hal-01769437.

Tsinaslanidis et Kugiumtzis 2014
Tsinaslanidis, Prodromos E. et Dimitris Kugiumtzis (2014). « A prediction scheme using perceptually important points and dynamic time warping ». In : *Expert Systems with Applications* 41.15, p. 6848–6860. ISSN : 0957-4174. DOI : https://doi.org/10.1016/j.eswa.2014.04.028. URL : http://www.sciencedirect.com/science/article/pii/S0957417414002516.

T.T.H. Phan et al. 2017
T.T.H. Phan et al. (2017). « Which DTW Method Applied to Marine Univariate Time Series Imputation ». In : *MTS/IEEE Oceans Conference - OCEANS'17 Aberdeen*. DOI : 10.1109/OCEANSE.2017.8084598.

Van Hoan, Huy et L.C. 2017
Van Hoan, M., D.T. Huy et Mai L.C. (2017). « Pattern Discovery in the Financial Time Series Based on Local Trend. » In : *Advances in Information and Communication Technology. ICTA 2016. Advances in Intelligent Systems and Computing. Springer, Cham. doi 10.1007/978-3-319-49073-1_.48* 538.

Wang et al. 2013
Wang, Xiaoyue et al. (2013). « Experimental comparison of representation methods and distance measures for time series data ». In : *Data Mining and Knowledge Discovery* 26.2, p. 275–309.

Xie et Wiltgen 2010
Xie, Ying et Bryan Wiltgen (2010). « Adaptive Feature Based Dynamic Time Warping ». In : *IJCSNS International Journal of Computer Science and Network Security* 10.1.

M. Zhang et al. 2018
Zhang, Min et al. (2018). « Spatiotemporal evolution of the chlorophyll a trend in the North Atlantic Ocean ». In : *Science of The Total Environment* 612, p. 1141–1148. ISSN : 0048-9697. DOI : 10.1016/j.scitotenv.2017.08.303.

Z. Zhang et Moore 2015
Zhang, Zhihua et John C. Moore (2015). « Chapters 1,2 and 6 ». In : *Mathematical and Physical Fundamentals of Climate Change*. Sous la dir. de Zhihua Zhang et John C. Moore. Boston : Elsevier, p. 1–47. ISBN : 978-0-12-800066-3. DOI : 10.1016/B978-0-12-800066-3.00001-2.

Chapitre 2

Comparaison de séries temporelles

« Le passé peut-il combler le présent ?
Vivre le présent différemment ?
Comment se comparer et avancer. »

EPC

Sommaire

2.1 Métriques pour la comparaison de signaux 22
2.1.1 Métriques statistiques 22
2.1.2 Métriques élastiques 25
2.1.3 Métriques multi-variables 27
2.2 Recherche de motifs pour l'imputation 28
2.2.1 Un point de vue sur les techniques d'imputation 28
2.2.2 Complétion par recherche d'un motif existant 29
2.3 Classification et Partitionnement de séries 41
2.3.1 Un point de vue sur les techniques de classification spectrale 41
2.3.2 Intégration de connaissances dans le processus de clustering. 42
2.3.3 Critère multi-coupe normalisé intégrant des contraintes de comparaison par paires. 43
2.4 Conclusions, Perspectives. 48

Ce chapitre est dédié à l'identification de sous-séquences de tailles fixées *a priori* et contient les points relatifs aux publications associées.

— La première section reprend les principales métriques utilisées pour comparer deux séquences.

— Le second concerne la recherche de motifs pour l'imputation de valeurs successives manquantes dans une série. Après avoir exposé un point de vue sur les techniques d'imputation, nous détaillons plusieurs approches proposées. (T.T.H. Phan, **E. Poisson-Caillault** et Bigand 2018 ; T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018 ; T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre et al. 2017 ; T.T.H Phan et al. 2017 ; T.T.H Phan, **E. Poisson Caillault** et Bigand 2019)

— La dernière section est dédiée à la comparaison de séries dans un but de partitionner les motifs similaires. (**Caillault** et al. 2009 ; P.A. Hébert, **E. Caillault Poisson** et Hamad 2011 ; G. Wacquet, **E. Poisson-Caillault** et Hebert 2013 ; G. Wacquet, **E. Caillault Poisson** et al. 2013 ; T.T.H. Phan, **E. Poisson Caillault** et Bigand 2016)

2.1 Métriques pour la comparaison de signaux

Quel que soit le cadre de l'identification, les notions de métrique et de représentation entre ces séquences restent à définir. Nous recenserons les métriques de comparaison de deux signaux de variables quantitatives discrètes les plus usitées dans la littérature. Et, nous démontrerons les pertinences et lacunes de chacune pour proposer une méthodologie basé sur un ensemble de critères à satisfaire.

2.1.1 Métriques statistiques

La table 2.1 reprend quelques métriques usuelles de comparaison entre deux séquences univariées notées q et r. Le signal r sera parfois considéré d'étendue ou moyenne non nulle et positive pour être exploitable pour certaines métriques. La colonne "Opt" renseigne si la métrique est à minimiser ou maximiser (respectivement min./max.). Les deux colonnes suivantes indique les niveaux d'interprétabilité de l'erreur, puis de fidélité à la forme entre les deux signaux (-/+/++ : - non interprétable, + pour interprétable facilement, ++ avec parfois des précisions sur les valeurs de la littérature considérée acceptable).

Le biais moyen – ou *Mean Bias*– (MB), appelé aussi erreur moyenne (moyenne des écarts – ou *Mean Error*– (ME)), est utilisé généralement comme mesure de sous ou sur-estimation entre un modèle q et une observation terrain r. La moyenne des écarts absolus – ou *Mean Absolute Error*– (MAE) et l'erreur quadratique moyenne des écarts – ou *Root Mean Square Error*– (RMSE) sont utilisés comme des descripteurs d'erreurs ; la moyenne des écarts absolus normalisée– ou *Normalized Mean Absolute Error*– (NMAE) et le biais moyen normalisé– ou *Normalized Mean Bias*– (NMB) sont des versions normalisées des indicateurs précédents permettant d'apporter des descripteurs plus facilement appréciables de cette erreur. Cependant ses mesures sont sensibles aux valeurs extrêmes ; un nombre faible d'écarts importants peut influencer significativement ces métriques.

Dans le même cadre, le biais fractionnel – ou *Fractional Bias*– (FB), la fraction des écarts-types – ou *Fractional Standard deviation*– (FS) (noté parfois aussi FD : Fractional Difference) ou sa variante fraction des écarts-types – ou *Fractional Standard Deviation*– (FSD) sont exploités pour compenser le manque d'interprétabilité de l'erreur et pouvoir composer directement avec des signaux de tailles différentes. Un modèle est jugé parfait quand FB et FS tendent vers zéro, acceptable lorsque $FB \leq 0,3$ et $FS \leq 0,5$.

La moyenne géométrique des écarts– ou *Geometric Mean Bias*– (BG) et la variance géométrique– ou *Geometric Mean Variance*– (VG) ont aussi l'avantage de proposer des limites d'acceptation du modèle.

Les similarités notées 'Sim' et 'Sim variante' dans le tableau permettent de juger la liaison de proximité et de forme entre deux signaux. Leurs équations sont définies à partir de deux signaux de même taille et basées sur un ratio entre l'aire des écarts point à point normalisé avec celle de l'original ou son étendue.

La covariance permet de quantifier la liaison entre deux variables quantitatives, notamment si les tendances de ces deux signaux par rapport à leurs espérances respectives sont proches. Plus sa valeur est élevée, plus la liaison est forte, mais elle ne permet pas de préjuger de la qualité contrairement à sa variante normalisée décrite ci-après. Le coefficient de détermination (R^2) basé sur le coefficient de corrélation empirique de Pearson permet de juger de la proportion de variance d'un signal linéairement expliquée par le second signal. Lorsqu'il est supérieur à $0,9$, la liaison est considérée forte.

Une autre approche est de considérer une enveloppe acceptable autour du signal à prédire et ainsi caractériser la liaison de forme entre les signaux. Le Facteur 2 (FA2) permet ainsi de calculer la fraction de données qui satisfait un tel critère. Un modèle est considéré comme parfait lorsque son FA2 est proche de 1 et acceptable lorsque $FA2 > 0,8$.

Métriques M(q,r)	Équations $\sum = \sum_{i=1}^{T}$	Opt	Domaine	Interprétabilité	
				Erreur	Forme
Conditions r et q de même taille $card(q) = card(r) = T$					
dMax	$max(\lvert q-r \rvert)$	min.	$\in \mathbb{R}^+$	-	-
MB=ME	$\overline{q-r}$	min.	$\in \mathbb{R}$	-	-
NMB=NME	$\frac{\overline{q-r}}{\overline{r}}$	min.	$\in \mathbb{R}$	+	-
MAE	$\overline{\lvert q-r \rvert}$	min.	$\in \mathbb{R}^+$	+	-
NMAE	$\frac{\overline{\lvert q-r \rvert}}{\overline{r}}$	min.	$\in \mathbb{R}^+$	-	-
NMAE variante	$\frac{\overline{\lvert q-r \rvert}}{\lVert max(r)-min(r) \rVert}$	min.	$\in \mathbb{R}^+$	+	-
RMSE	$\sqrt{\frac{1}{T}\sum(q_i-r_i)^2}$	min.	$\in \mathbb{R}^+$	-	
NMSE	$\frac{\overline{(q-r)}^2}{\overline{q}\times\overline{r}}$	min.	$\in \mathbb{R}^+$	-	
NMSE variante	$\sqrt{\frac{\sum(q-r)^2}{\sum r^2}}$	min.	$\in \mathbb{R}^+$	-	
Sim	$\frac{1}{T}\sum\frac{1}{1+(q_i-r_i)^2}$	max.	$\in [0,1]$	+	+
Sim variante	$\frac{1}{T}\sum\frac{1}{1+\frac{\lvert q_i-r_i \rvert}{\lvert max(r)-min(r) \rvert}}$	max.	$\in [0,1]$	+	+
GMV=VG	$\exp(\overline{(\ln(q)-\ln(r))^2})$	min.	$\in \mathbb{R}^+$	$+\ 0,75 \leq VG \leq 1,25$	+
FA2	$\frac{1}{T}\text{card}(0,5 \leq q_i/r_i \leq 2)$	max.	$\in [0,1]$	+ FA2 $> 0,8$	+
R^2 Pearson	$(\frac{cov(q,r)}{\sigma_q\times\sigma_r})^2$	max.	$\in [0,1]$	++ $R^2 \geq 0,9$ p-value	–
$card(r) = Tr$ et $card(q) = T_q$					
Biais	$\overline{q}-\overline{r}$	min.	$\in \mathbb{R}$	-	-
FB	$2\times\frac{\overline{q}-\overline{r}}{\overline{q}+\overline{r}}$	min.	$\in [-2,2]$	++ $\lvert FB \rvert \leq 0,3$	-
GMB=BG	$\exp(\overline{\ln(q)}-\overline{\ln(r)})$	min.	$\in \mathbb{R}^+$	$+\ 0,75 \leq BG \leq 1,25$	+
FS	$2\times\frac{\lvert(\sigma_q)^2-(\sigma_r)^2\rvert}{(\sigma_q)^2+(\sigma_r)^2}$	min.	$\in [0,2]$	++ $FS \leq 0,5$	-
FSD	$2\times\lvert\frac{\sigma_q-\sigma_r}{\sigma_q+\sigma_r}\rvert$	min.	$\in [0,2]$	++ $FSD \leq 0,5$	-

TABLEAU 2.1 – Métriques classiques de comparaison entre deux sous-séquences ou séquences univariées q et r. Les signes -/+/++ indiquent le niveau d'interprétabilité : (-) du non quantifié donc peu interprétable à (++) pleinement avec des bornes.

Illustrations numériques.

Pour comprendre la philosophie des approches et algorithmes proposés par la suite, une étude quantitative et illustratrice de comparaison de signaux élémentaires a été proposée dans la thèse de Hong Phan (chapitres 1 et 2, PHAN 2018) et complétée ici sur les métriques

fréquemment utilisées dans la littérature pour comparer un signal prédit à un signal observé ou deux signaux entre eux, reportés dans la table 2.1.

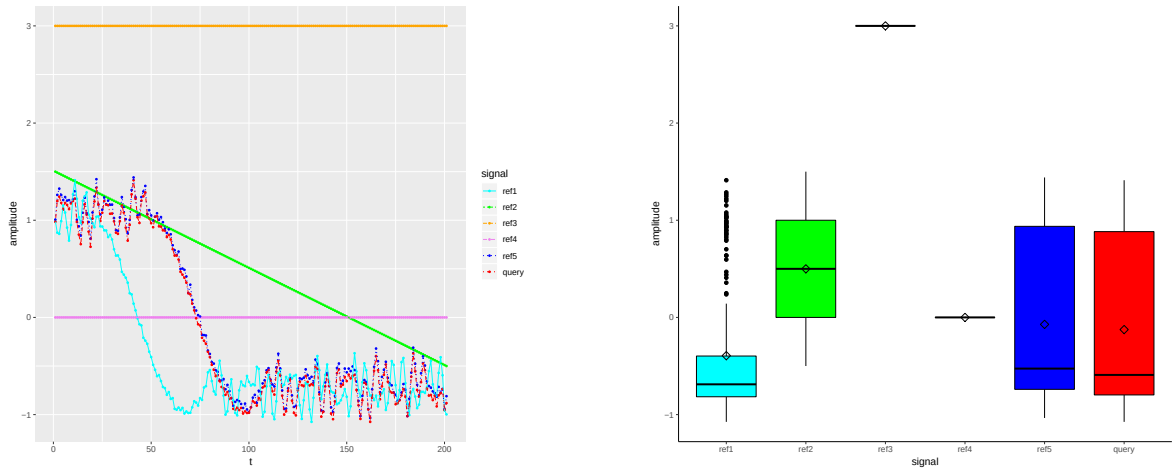


FIGURE 2.1 – Signaux élémentaires artificiels et leurs statistiques de base.

La figure 2.1 illustre six signaux élémentaires générés. Ces signaux sont disponibles dans le package R DTWBI sur le site du CRAN et développé dans le cadre du CPER MARCO. Ils ont été choisis de la manière suivante : le signal 'query' borné entre -1 et 1 est obtenu à partir du signal 'ref5' de mêmes bornes auquel il a été rajouté un bruit gaussien borné entre 0 et 0,1. Les signaux 'ref1' et 'ref5' sont générés à partir du même signal, avec un déphasage temporel. Les signaux 'ref2', 'ref3' et 'ref4 'sont trois signaux linéaires. 'ref4' est d'amplitude proche de la moyenne des signaux 'ref5' et 'query', 'ref 3' avec un offset supérieur à l'étendue de ces signaux. Et, ref2 est un signal décroissant de tendance générale proche de 'ref5'.

	query	ref1	ref2	ref3	ref4	ref5
distMax	0,00	1,64	1,53	4,07	1,41	0,10
ME	0,00	0,27	-0,63	-3,13	-0,13	-0,05
NME	0,00	-0,68	-1,25	-1,04	-	0,75
MAE	0,00	0,42	0,64	3,13	0,80	0,05
NMAE	0,00	1,06	1,28	1,04	-	0,75
NMAEv	0,00	0,17	0,32	-	-	0,02
RMSE	0,00	0,04	0,05	0,23	0,06	0,00
NMSE	0,00	7,87	-9,62	-27,88	-	0,41
NMSEv	0,00	0,74	0,92	3,84	1,00	0,07
sim	1,00	0,76	0,57	0,02	0,40	0,99
simv	1,00	0,87	0,78	0,00	0,00	0,98
VG	1,00	14,71	12,71	137,54	-	**1,02**
FA2	1,00	0,78	0,41	0,00	0,00	**0,99**
R2,cor	1,00	0,55	0,72	0,00	0,00	**1,00**
Bias	0,00	0,27	-0,63	-3,13	-0,13	-0,05
FB	0,00	-1,04	-3,33	-2,17	2,00	0,55
BG	1,00	1,33	**0,78**	0,18	-	**0,98**
FS	0,00	**0,41**	0,69	2,00	2,00	**-0,00**
FSD	0,00	**0,21**	**0,36**	2,00	2,00	**0,00**
1 critère	oui	oui	oui	non	non	oui
tout critère	oui	non	non	non	non	non

TABLEAU 2.2 – Étude des métriques statistiques entre le signal query et l'ensemble des signaux élémentaires (en gras les valeurs respectant les bornes acceptables du critère)

Les valeurs des métriques entre ces signaux élémentaires sont reportées dans le tableau 2.2. Les premières métriques ne permettent pas de discriminer fortement le signal le plus proche du signal 'query' parmi l'ensemble des i signaux de références notés 'refi'. Les similarités avec une valeur inférieure à $0,5$ permettent d'écarter les signaux plats 'ref3' et 'ref4'. En gras, sont mises en avant les valeurs respectant le critère borné donné au tableau 2.1. 'ref1', 'ref2' et 'ref5' satisfont au moins un des critères. Le biais géométrique considère le signal linéaire de tendance proche 'ref2' comme un candidat valable, proche de 'query', il ne permet pas d'interpréter localement le respect de la forme du signal de même les métriques FS/FSD. Bien que le signal 'ref5' soit le plus proche de la requête, il ne satisfait pas tous les critères valués des métriques.
Sur ce choix de représentation à savoir de partir des signaux bruts, ces métriques usuelles ne permettent pas de répondre à l'ensemble des critères d'erreur ou de forme acceptable et sont très sensibles à des décalages temporels.

2.1.2 Métriques élastiques

Deux signaux avec un décalage temporel ou d'amplitude peuvent être considérés comme similaire si la forme générale est respectée. Tel est le cas des signaux query, 'ref1' et'ref5'. Dans ce cas, il convient d'opérer une déformation élastique pour être tolérant à ce décalage et chercher le meilleur appariement de ces signaux. Les signaux alignés devenant ainsi de même taille, toutes les métriques précédentes pourront être appliquées et seront qualifiées de métriques élastiques auquel on ajoutera comme métrique le coût de déformation.

DTW - Dynamic Time Warping
L'algorithme DTW (Sakoe et Chiba 1978) consiste à calculer la matrice de coût de déformation élastique entre deux signaux et utiliser la programmation dynamique pour extraire le meilleur chemin d'appariement illustré figure 2.2 dans un cadre général.

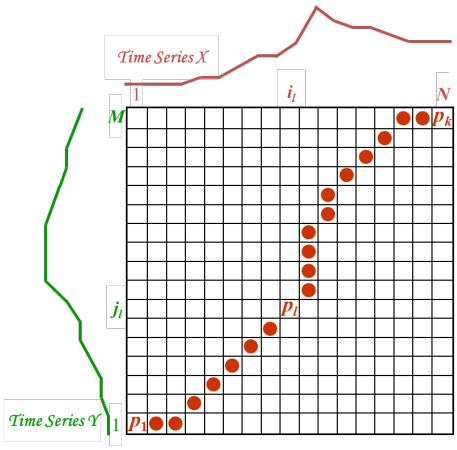


Figure 2.2 – Matrice de coût et chemin optimal d'appariement, tirée de (GenTXWarper p.d.)

Cette matrice $DTW(q,r) = \{dist(i,j)\}$ est définie par l'équation (2.1) où d représente la distance euclidienne, et par les conditions de continuité, monotonie et d'appariement forcé des points initiaux et finaux de chaque signal. La figure 2.3 illustre l'appariement entre les deux signaux 'query' et 'ref1' et à droite leurs déformations avec une tolérance de décalage de 10 % de la taille totale du signal de référence.

$$dist(i,j) = d(q_i, r_j) + min\{dist(i-1,j-1), dist(i-1,j), dist(i,j-1)\} \tag{2.1}$$

De l'algorithme originel de Sakoe et Chiba ont été déclinés différents types d'appariement, linéaire, contraint, ..., basés sur un calcul de distance prenant en compte des spécificités lo-

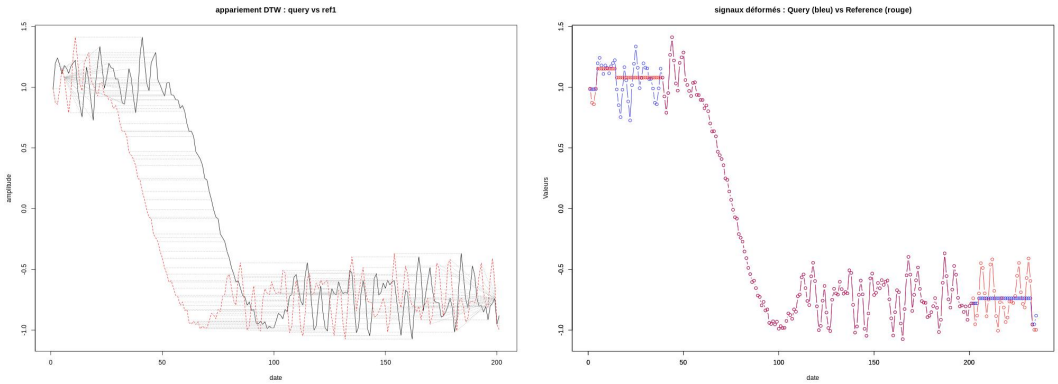


FIGURE 2.3 – Appariement et déformation DTW pour les signaux 'query' et 'ref1'

cales et/ou générales. Nous reprenons ici ces derniers.

DDTW - Derivative DTW.
E. J. KEOGH et M. J. PAZZANI 2001 remplace les signaux en entrée de l'algorithme par leurs dérivées locales. Soit (r,q) les deux signaux à comparer, l'algorithme originel sera appliqué sur (Dr, Dq) tel que pour un signal $x = \{x_i, i \in \{1, \cdot, n\}\}$, sa dérivée locale Dx est définie par :

$$Dx = \frac{(x_i - x_{i-1}) + \frac{x_{i+1} - x_{i-1}}{2}}{2}, 1 < i < N \tag{2.2}$$

Le coût d'appariement noté $DDTW(q, r)$ correspond alors au coût $DTW(Dq, Dr)$.

AFBDTW - Adaptive Feature Based DTW.
Les tendances et formes générales des signaux ne sont pas forcément respectées après un appariement selon DTW ou DDTW. XIE et WILTGEN 2010 ont proposé une modification de la matrice de coût basée à la fois sur écarts entre des caractéristiques locales telle la dérivée f_{local} et des écarts entre les caractéristiques globales f_{global} et une distance de manhattan au lieu de la norme L2 initiale. Le coût d'appariement d'une paire de points $AFBDTW(q_i, r_j) = AFBDTW(i, j)$est alors défini par les équations suivantes

$$AFBDTW(i, j) = w_1 \times dmanhattan_{local}(q_i, r_j) + w_2 \times dmanhattan_{global}(q_i, r_j) \tag{2.3}$$

$$f_{local}(x_i) = ((f_{local}(x_i))_1, (f_{local}(x_i))_2) = (x_i - x_{i-1}, x_i - x_{i+1}) \tag{2.4}$$

$$f_{global}(x_i) = ((f_{global}(x_i))_1, (f_{global}(x_i))_2) = (x_i - \sum_{k=1}^{i-1} \frac{x_k}{i-1}, x_i - \sum_{k=i+1}^{N} \frac{x_k}{N-i}) \tag{2.5}$$

w1 et w2 sont des paramètres de pondération à calibrer fonction de l'application et des contraintes souhaitées aux formes locales ou aux tendances globales.

DTW-D : Normalisation du critère DTW
CHEN, HU et al. 2013 ont proposé une normalisation du coût DTW par la distance euclidienne (ED) entre les signaux. Pour une application particulière de classification de séries, cette mesure DTW-D définie par l'équation 2.6 a permis de mieux discriminer des signaux de faibles amplitudes.

$$DTW - D(q, r) = \frac{DTW(q, r)}{ED(q, r) + \epsilon} \tag{2.6}$$

Expérimentations numériques
Le tableau 2.3 transcrit les coûts d'appariement entre le signal 'query' et les 5 signaux de référence. La métrique ED peut être interprétée comme une métrique d'élasticité nulle : ici les signaux étant de même taille initialement, l'alignement est réalisé date à date entre les

deux signaux. En gras, sont mis en avant les valeurs pour lesquelles le coût de déformation est le plus faible. Ainsi, nous pouvons remarquer que le décalage temporel du signal 'ref1' par rapport à 'query' n'est pas ou peu pénalisé par rapport au signal 'ref5' sans décalage et légèrement bruité.

query vs	DTW	DDTW	AFBDTW	ED	DTW-D
ref1	**0,02**	**0,02**	0,14	8,83	**0,002**
ref2	0,11	0,04	0,32	11,00	0,01
ref3	2,35	0,06	0,56	45,85	0,05
ref4	0,41	0,05	0,56	11,95	0,03
ref5	0,04	**0,02**	**0,05**	**0,86**	0,04

TABLEAU 2.3 – Coûts d'appariement entre 'query' et signaux de références selon les métriques élastiques choisies (en gras les valeurs respectant les bornes acceptables du critère).

Les métriques ayant un critère d'interprétabilité sont reprises dans le tableau 2.4, elles sont calculées à partir des signaux alignés.

	VG	FA2	R2,cor	FB	BG	FS	FSD	tout critère
ref1-DTW	**1,04**	**1,00**	**0,99**	**0,04**	**1,01**	**-0,01**	**0,00**	**oui**
ref1-DDTW	**1,08**	**1,00**	**0,98**	**-0,05**	**0,97**	**0,14**	**0,07**	**oui**
ref1-AFBTDTW	**1,06**	**1,00**	**0,99**	**-0,03**	**1,02**	**0,05**	**0,03**	**oui**
ref2-DTW	g,	**0,96**	**0,96**	-0,47	89,37	**0,23**	**0,11**	-
ref2-DDTW	83,98	0,41	0,33	-8,43	1,57	-0,50	**0,25**	-
ref2-AFBTDTW	3,63	0,42	**0,92**	-2,84	1,35	**0,05**	**0,03**	-
ref3-DTW	12,08	0,00	0,00	-1,30	0,39	2,00	2,00	-
ref3-DDTW	g,	0,00	0,00	-2,87	0,09	2,00	2,00	-
ref3-AFBTDTW	g,	0,00	0,00	-2,70	0,14	2,00	2,00	-
ref4-DTW	g,	0,00	0,00	2,00	g,	2,00	2,00	-
ref4-DDTW	g,	0,00	0,00	2,00	g,	2,00	2,00	-
ref4-AFBTDTW	g,	0,00	0,00	2,00	g,	2,00	2,00	-
ref5-DTW	**1,04**	**0,99**	**1,00**	**0,19**	**1,07**	**-0,00**	**0,00**	**oui**
ref5-DDTW	**1,05**	**0,94**	**0,99**	0,68	**0,99**	**-0,05**	**0,03**	-
ref5-AFBTDTW	**1,02**	**0,99**	**1,00**	0,52	**0,98**	**-0,01**	**0,00**	-

TABLEAU 2.4 – Métriques statistiques sur les signaux alignés après une recherche d'appariement avec le signal 'query', (en gras les valeurs pour lesquels les critères sont respectés)

En lien avec les conclusions du tableau précédent, le décalage du signal 'ref1' est toléré et permet de statuer le signal comme proche de 'query' quel que soit le type d'appariement élastique (certes avec une tolérance d'élasticité de 10 %). Le coefficient de détermination et le FA2 sont trop souples : ils satisfont leur critère d'interprétabilité pour un signal linéaire de tendance proche. Pour discriminer ces cas, les algorithmes d'alignement plus coûteux en calcul tels DDTW ou AFBDTW seront à privilégier. L'appariement DTW reste le plus utilisé dans la littérature. Au regard de ces chiffres, il est aussi tout à fait satisfaisant en utilisant l'ensemble de ces critères et non uniquement les critères (R^2, FA2, FS, FSD) souvent usités.

2.1.3 Métriques multi-variables

Lorsque les segments à comparer sont composés de plusieurs variables, les normes L_1, L_2 et L_∞ sont alors utilisées.

Les fonctions de coût DTW, DDTW et AFBDTW sont des mesures relatives moyennées dépendant de l'intensité des deux signaux. Dans (**E. Caillault**, HEBERT et WACQUET 2009), nous avons proposé une mesure de dissimilarité bornée entre 0 et 1 pour des signaux à valeurs

positives (transposable au cadre des réels). Cette dissimilarité ds est basée sur un ratio entre l'écart local et la valeur maximale d'un des deux signaux à comparer. Avec d la distance euclidienne, elle est définie par :

$$ds(ra_i, qa_j) = \frac{d(ra_i, qa_j)}{max(d(ra_i, 0), d(qa_j, 0))} \tag{2.7}$$

Dans le cas de signaux multivariés $R = \{r_{ik}\}$ et $Q = \{q_{jk}\}$ avec i,j les indices temporels et k l'indice de la variable, l'alignement est contraint simultanément sur l'ensemble des signaux univariés en préférant une distance de Manhattan pour accumuler l'ensemble des distorsions sur chaque variable.

$$ds(ra_i, qa_j) = \frac{1}{n_c} ds(ra_{i_k}, qa_{j_k}) \tag{2.8}$$

Deux cadres d'application de ces métriques seront présentés par la suite : l'une destinée à l'imputation de séries à valeurs manquantes, la seconde au partitionnement non supervisé de séries temporelles. Nous utiliserons le terme efficace lorsqu'une méthode remplit l'ensemble des critères bornés du tableau 2.1.

2.2 Recherche de motifs pour l'imputation

Les méthodes d'analyses exploratoires et de classification des séries requièrent des données complètes, c'est-à-dire sans valeurs manquantes pour ne pas biaiser l'interprétation ou simplement pour pouvoir les appliquer. Ces absences de valeurs sont généralement liées aux défaillances capteurs, aux difficultés de mise en oeuvre de la maintenance ou encore aux problèmes de transmission des données ou encore provoquées par des valeurs aberrantes supprimées. Elles sont fréquentes et parfois sur des durées importantes, notamment dans le cadre marin (Ceong, Kim et Park 2012).
Nous définirons ici une sous-séquence $Y_{[a,b]}$ d'une série Y de N_t observations comme le sous-ensemble des valeurs des variables entre les positions a et b avec $a \geq 1$ et $b \leq N_t$, deux valeurs naturelles connues.

Definition 2.2.1. La sous-séquence $Y_{[a,b]}$ d'une série Y de N_t observations de C variables est définie par : $Y_{[a,b]} = \{Y(c,t),\ t \in [a,b] \subset [1, N_t], c \in [1, \ldots, C]\}$

2.2.1 Un point de vue sur les techniques d'imputation

Un grand nombre de techniques ont été explorées pour compléter des séries temporelles multivariées. Lorsque les signaux ou leurs caractéristiques sont fortement corrélés, une valeur manquante à l'instant m d'un signal y_c peut alors être estimée à partir des valeurs des signaux disponibles tel que $y_{c,m} = f(y_{i,m} | i \neq c)$. Les techniques de régression à partir d'une base d'apprentissage constitué de l'ensemble des observations complètes ont été particulièrement appliquées à cette tâche avec des modèles discriminants tels les K-plus-proche-voisins (Kévin Rousseeuw 2014 ; Liao et al. 2014 ; Rahman et al. 2015) et les forêts aléatoires (RF - Random Forest) (Daniel J. Stekhoven et Bühlmann 2012), ou avec des modèles génératifs basés sur des hypothèses de distribution normale multivariée (Schafer 1997). Les valeurs imputées sont alors générées par des méthodes de Monte-Carlo par chaînes de Markov (MCMC - Markov Chain Monte Carlo). Lorsqu'aucune distribution conjointe n'est trouvée, des distributions conditionnelles viennent relâcher le problème : approches MICE (Multiple Imputation Chained Equations) (Van Buuren, Boshuizen, Knook et al. 1999 ; Raghunathan et Siscovick 1996 ; Raghunathan, Lepkowski et al. 2001 ; Stuart et al. 2009 ; Royston 2007 ; Joseph et al. 2009 ; Lee et Carlin 2010 ; Spratt et al. 2010 ; Gelman

et al. 2015 ; Deng et al. 2016) ou une combinaison des deux modèles MICE et RF (Shah et al. 2014).

Lorsque les signaux sont peu ou non corrélés entre eux, une autre approche consiste à étudier les signaux indépendamment. Les approches peuvent être découpées selon le niveau de stationnarité du signal. Lorsqu'un signal est fortement stationnaire, les modèles auto-régressifs et moyenne mobile permettent de prédire la valeur manquante $y_{c,m} = f(y_{c,t}|t \leq m)$. Les techniques élémentaires telles la copie de la dernière valeur présente (na.locf last observation carried forward), complétion par moyenne (Allison 2001 ; Bishop 2006) ou des interpolations par moyenne ou spline (Zeileis et Grothendieck 2005) permettent de compléter des séries non ou faiblement stationnaires avec des trous isolés ou de taille élémentaire (unitaire ou [i,m] faible vis-à-vis de la dynamique du signal). Les approches par modélisation ARIMA (AutoRegressive Integrated Moving Average) ou SARIMA (Seasonal-ARIMA) sont aussi très en vogue dans la littérature lorsque le signal est non stationnaire. Les modèles ARIMA permettent de modéliser des séries temporelles qui présentent une tendance polynomiale et SARIMA une saisonnalité (Walter.O et al. 2013).

Dans le cas de notre sujet d'étude des efflorescences phytoplanctoniques, les signaux issus de la station Marel Carnot en rade de Boulogne-sur-mer (illustrés chapitre 1 et 3, détaillés dans Kévin Rousseeuw 2014 ; Alain Lefebvre 2015) présentent une variabilité saisonnière et sans tendance générale ; les méthodes précitées ne sont alors pas applicables. Plusieurs études ont été menées dans un cadre restrictif de valeurs manquantes isolées ou de sous-séquences de taille réduite (Moritz, Sardá et al. 2015 ; Junninen et al. 2004). Mais Avec un cadre élargi quant aux types de signaux et/ou des séquences continues de valeurs manquantes de tailles variables, nous avons montré que ces approches ne permettent pas une reconstruction efficace des signaux ou de leur dynamique (T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre et al. 2017 ; T.T.H Phan et al. 2017).

Pour la suite de ce document, nous qualifions la notion de trou par les définitions suivantes.

Definition 2.2.2. Un T-trou (T-gap) est une sous-séquence $Y_{[a,b=a+T-1]}$ de T valeurs manquantes consécutives (NA : Not Available) pour au moins une des variables :
$\forall t \in [a,b]\ \exists c,\ Y(c,t) = \text{NA}, Y(c,a-1) \neq \text{NA},\ Y(c,b+1) \neq \text{NA}.$

Definition 2.2.3. Un $T = 1$-trou appelé aussi **trou isolé est une observation $Y_{[t]}$ avec une seule 1 valeur manquante pour au moins une des variables :**
$\exists (c,t),\ Y(c,t) = \mathbf{NA}, Y(c,t-1) \neq \mathbf{NA},\ Y(c,t+1) \neq \mathbf{NA}.$

Pour compléter des trous de taille dite large vis-à-vis de la dynamique du signal/processus, une hypothèse, certes forte mais assez réaliste dans les applications, est de considérer une récurrence des phénomènes observés :

— la sous-séquence $Y_{[m-T,m-1]}$ précédant un trou de taille T débutant à l'instant m est déjà existante dans la base de données collectées et,

— cette sous-séquence entraîne la même dynamique particulière qu'il sera possible de répliquer ou adapter dans le trou $Y_{[m,m+T-1]}$.

Ainsi l'hypothèse de dynamique locale déjà existante sur le signal lui-même ou l'ensemble des signaux permet d'estimer la valeur manquante $y_{c,m} = f(y_{c,k}|k$ indéterminé).

2.2.2 Complétion par recherche d'un motif existant

La figure 2.4 illustre la technique de complétion d'un T-Trou par recherche d'une récurrence du phénomène observé précédant ce T-trou. Il est supposé que ce phénomène entraînera les mêmes conséquences et donc la même forme du signal.

Cette approche est décomposée en 4 étapes :

1. construire une fenêtre $Q_{[m-T,m-1]}$ précédant le T-Trou $Y_{[m,m+T-1]}$;

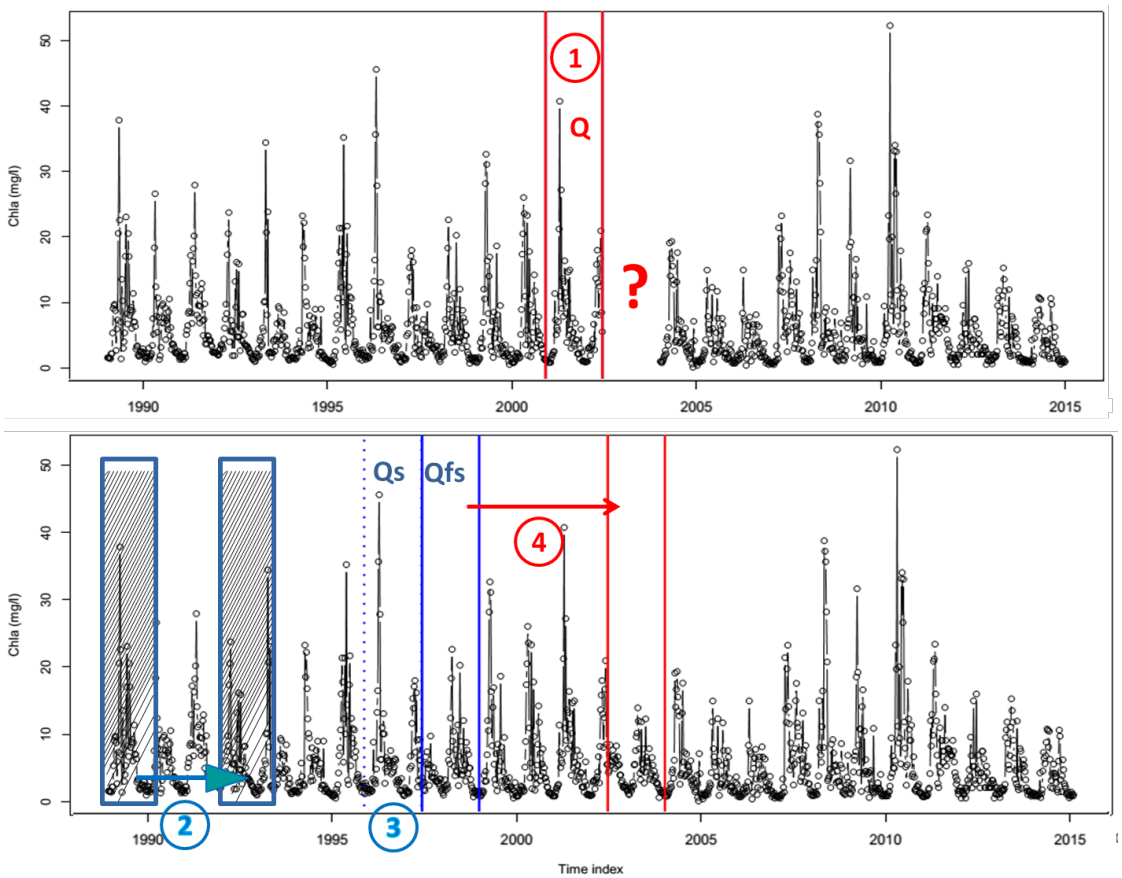


FIGURE 2.4 – Illustration de la recherche d'un motif identique à la séquence Q précédant un T-Trou signifié ici par le symbole ?. 1- Construction de la séquence recherchée Q, 2-Comparaison des séquences proches par fenêtre glissante Qf, 3-Sélection de la fenêtre le plus similaire Qfs, 4- Imputation du T-Trou.

2. comparer une fenêtre R sans valeur manquante d'une série de référence comme par exemple la série passée et/ou future à laquelle appartient le trou ;
3. Par fenêtre glissante, on identifiera la ou les plus proches notées $R = Qs$ selon un critère donné.
4. Les fenêtres immédiatement consécutives à celles-ci - si elles existent - seront retenues et notées Qfs pour compléter le T-Trou par recopie de la plus similaire ou fusion des solutions.

DTWBI - Dynamic Time Warping Based Imputation

DTWBI est le nom à la fois de notre premier algorithme de complétion et du package R déployé sur le dépôt CRAN (https://cran.r-project.org/web/packages/DTWBI/index.html). Il se limite à la complétion d'un T-trou d'une série univariée. Il procède par recopie de la séquence consécutive de la fenêtre la plus similaire selon le critère de déformation DDTW. Ce critère étant très coûteux, il a été proposé de sélectionner un nombre réduit de fenêtres admissibles et calculer ce critère DDTW si elles respectent un seuil de proximité basé cosinus entre neuf caractéristiques globales relatifs à la forme du signal (T.T.H. PHAN, **E. Poisson Caillault** et BIGAND 2016).

Expérimentations numériques.
Cet algorithme a dans un premier temps été testé sur 9 jeux de données dont les caractéristiques sont décrites brièvement dans le tableau suivant 2.5. Les 5 premiers sont très utilisés dans la littérature pour éprouver les méthodes d'imputation. La sixième série a été obtenue

Algorithme 2 Algorithme d'imputation DTWBI d'une série monovariée.

Require: $x = \{x_1, x_2, \ldots, x_N\}$: série trouée
t : position initiale du trou)
T : taille du Trou
θ_cos : seuil de proximité en cosinus (≤ 1)
$step_threshold$: incrément de fenêtre glissante pour la recherche de seuil
$step_sim_win$: incrément de fenêtre glissante pour la sélection de fénêtre admissible $<<$ $step_threshold$.

Lop : liste de positions des fenêtres admissibles
LCost : liste des coûts DTW
$y = x$ série imputée renvoyée
Step 1 : $Dx = DDTW(x)$
Step 2 : Construire $Q = Dx[t - T : t - 1]$
Step 3 : Contruire une base de référence SDB non trouée : $SDB = Dx[1 : t - 2T]$ non trouée
Step 4 : Recherche d'un seuil DTW
$i \leftarrow 1$; $Lcost \leftarrow NULL$
while $i <= length(SDB)$ **do**
  $k \leftarrow i + T - 1$
  créer la fenêtre glissante R : $R(i) = SDB[i : k]$
  calculer les caractéristiques globales de Q et $R(i)$: gfQ, gfR
  calculer : $cos = cos(Q, R) = cosine(gfQ, gfR)$
  if $cos \geq \theta_cos$ **then**
    calculer coût DTW : $cost = DTW_cost(Q, R(i))$
    mémoriser le coût $cost$ dans LCost
  end if
  $i \leftarrow i + step_threshold$
end while
$threshold = min\{LCosts\}$
Step 5 : Rechercher des fenêtres admissibles
$i \leftarrow 1$; $Lop \leftarrow NULL$; $LCost \leftarrow NULL$
while $i < length(SDB)$ **do**
  $k \leftarrow i + T - 1$
  créer la fenêtre glissante R : $R(i) = SDB[i : k]$
  calculer les caractéristiques globales de Q et $R(i)$: gfQ, gfR
  calculer : $cos = cos(Q, R) = cosine(gfQ, gfR)$
  if $cos \geq \theta_cos$ **then**
    calculer coût DTW : $cost = DTW_cost(Q, R(i))$
    if $cost < threshold$ **then**
      enregistrer la position de $R(i)$ dans Lop
      mémoriser le coût $cost$ dans LCost
    end if
  end if
  $i \leftarrow i + step_sim_win$
end while
Step 6 : Remplir le T-Trou
$i = argmin(LCost)$; $p = Lop(i)$
$y_{[t,t+T-1]} = x_{[p,p+2T-1]}$
return y - série complétée

à partir de l'équation de Mackey-Glass est choisie pour étudier le comportement de ces approches face à un processus chaotique. Les suivantes concernent nos projets : la septième pour une évaluation climatologique données fournies par l'université de Hanoi et la dernière pour l'évaluation des masses d'eau au large de Boulogne-sur-Mer avec une série issue de MAREL-Carnot. Le dernier, la concentration d'ammonium (NH4+) issue du challenge industriel GECCO 2014 pour comparaison avec les travaux dans (Moritz et Bartz-Beielstein 2017).

n°	Nom	Longueur	Tend.	Sais.	Relevé	CC-10 %.	provenance
1	Air passenger	144	oui	oui	mensuel	0,94	R-package TSA
2	Beersales	192	oui	oui	mensuel	0,87	R-package TSA
3	Google	521			journalier	0,5	R-package TSA
4	SP	168	oui	oui	trimestriel	0,67	R-package TSA
5	CO2 concentrations	160	oui	oui	mensuel	0,98	Thoning, Tans et Komhyr 1989
6	Mackey-Glass chaotic	1 201			-	0,99	-
7	Phu Lien temperature	648		oui	mensuel	0,7	Univ. Hanoï
8	Water level	131 472		oui	20 minutes	1	MAREL-Carnot
9	tsNH4Complete	4 552	oui	oui	10 minutes		Moritz et Bartz-Beielstein 2017

Tableau 2.5 – Caractéristiques de jeux de données. (Tend. : tendance, Sais. : saisonnalité, CC-10 % : maximum de cross-corrélation entre le signal et des T-Trou de taille égale à 10 % du signal.

DTWBI a été comparé à plusieurs approches d'imputation, de la plus élémentaire adaptés au trou isolé jusqu'à des estimateurs basés sur des décomposition tendance-cycle ou des modèles aux composantes inobservables :

— na.locf (last observation carried forward, zoo R-package) : algorithme de remplacement par la dernière valeur observée avant le trou.

— na.aggregate (R-package zoo) : complétion des données manquantes par une moyenne totale ou imposée (mensuelle, journalière, ...).

— na.spline (R-package zoo) : imputation par interpolation polynomiale.

— na.interp (R-package forecast) : interpolation linéaire pour les séries sans saisonnalité ou par lissage des données corrigées de la saisonnalité avec une régression linéaire pondérée. Cette méthode est très adaptée aux séries ave des cycles très marqués.

— na.approx (zoo R-package zoo) : identique à la précédente sans correction de la saisonnalité.

— na.kalman (R-package imputeTS) : lissage selon un modèle espace-état de type ARIMA dont les paramètres sont estimés par un filtre de Kalman (Moritz et Bartz-Beielstein 2017).

Le tableau 2.6 reprend une fraction des études faites dans T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre et al. 2017 ; Phan 2018 ou `http://mawenzi.univ-littoral.fr/DTWBI/example/` où je renvoie le lecteur pour les protocoles d'expérimentations complets. Ces indicateurs de qualité de l'imputation sont moyennés à partir des résultats issus de la création de 10 à 50 T-Trous d'une taille T égale à 10 % de la taille totale de chaque série (le nombre de trous étant lié à la taille du jeu de données). L'indice de cross-corrélation donné au tableau 2.6 est calculé entre les fenêtres précédant un T-Trou et toutes les fenêtres glissantes de la série réalisées pour comparer leur proximité de forme et valeurs. Il est à mettre en relation avec ces résultats. En effet, plus cet indice est élevé, plus l'hypothèse de trouver une récurrence du phénomène observé est forte et plus il sera aisé de compléter les valeurs manquantes de la série par des méthodes élémentaires. Cet indice est un bon indicateur pour savoir si la méthode DTWBI a un intérêt pour l'imputation puisqu'il est lui-même basé sur ce critère de récurrence. Ainsi, les T-trous générés pour les séries Google et SP avec des coefficients de cross-corrélation inférieurs ou égaux à 50 % (resp. $0,5$ et $0,67$ pour une taille de trou de 10 %, et $0,4$ - $0,65$ pour une taille de 15 %) ont des indicateurs de similarité avec la vérité terrain moins percutants mais tout à fait proches des autres techniques comme l'interpolation

linéaire ou la moyenne (na.aggregate). Pour l'ensemble des autres jeux, la méthode proposée DTWBI obtient des résultats intéressants quel que soit l'indicateur.

Dans un cadre de prédiction météorologique, l'algorithme DTBWI a été appliqué à la prédiction du futur sur le même principe que l'imputation de T-Trou : considérant que nous sommes aujourd'hui à l'instant t, il manque la connaissance de T valeurs suivantes. Selon la taille de la prédiction souhaitée court terme ou long terme, la taille T de la construction de la requête Q sera adaptée. Ceci a été comparé avec des approches naïves recopiant la valeur précédente à la même saisonnalité (seasonal-naive),des prédicteurs par processus non stationnaires comportant des saisonnalités (SARIMA, Shumway et Stoffer 2017) ou par une modélisation statistique bayésienne (BSTS - Bayesian Structural Time Series, Scott et Varian 2014) et enfin par réseau de neurones à une couche cachée (nnetar du package-R forecast (Hyndman et Khandakar 2008)). Des expérimentations rapportées dans (T.T.H. Phan, **E. Poisson-Caillault** et Bigand 2018), la technique de prédiction neuronale apparaît la plus performante sur des critères d'écart d'erreur notamment pour des données mensuelles. La méthode DTBWI, non loin de ce dernier en terme de critères d'écart, permet de mieux respecter la forme du signal.

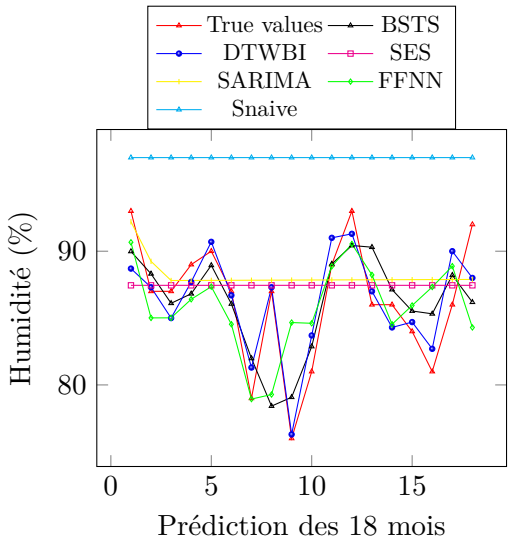


Figure 2.5 – Prédictions sur 18 mois du taux d'humidité de Phu Lien selon différentes techniques comparées aux valeurs observées.

Sur la figure 2.5, le signal prédit DTWBI (points bleus) pour 18 prochains mois suit plus fidèlement la réalité du phénomène d'humidité mesurée à Phu Lien (tracé rouge avec triangle). Notamment la différence est notable avec le signal prédit par FFNN (en vert, symbole losange) entre le sixième et dixième mois et après le 14ème mois.

DTWUMI - DTW-based Imputation for Uncorrelated Multivariate Series

DTWUMI est une extension naturelle de l'algorithme DTWUBI aux cas de séries multivariées où l'on recherchera une récurrence du phénomène observé sur l'ensemble des signaux variables constituant la série $X = \{x_{ik} = x_k(t = i)\}$. La requête Q initialement un segment dans DTWBI devient une matrice Q de taille $T \times C$ illustrée à la figure 2.6. Pour rappel, T est la taille du T-trou sur une ou plusieurs variables et C le nombre de variables.

Ce processus d'imputation impose le recours à des trapèzes (figure basse 2.7) qui sont insérés dans la série initiale (figure haute) pour la recherche de la fenêtre la plus similaire. Ces trapèzes servent à combler la série de référence qui doit être sans valeur manquante. Ils sont définis par l'équation donné à la figure 2.7 avec une homothétie liée à l'étendue du signal comblé et la taille de chaque trou. Contrairement aux approches de la littérature

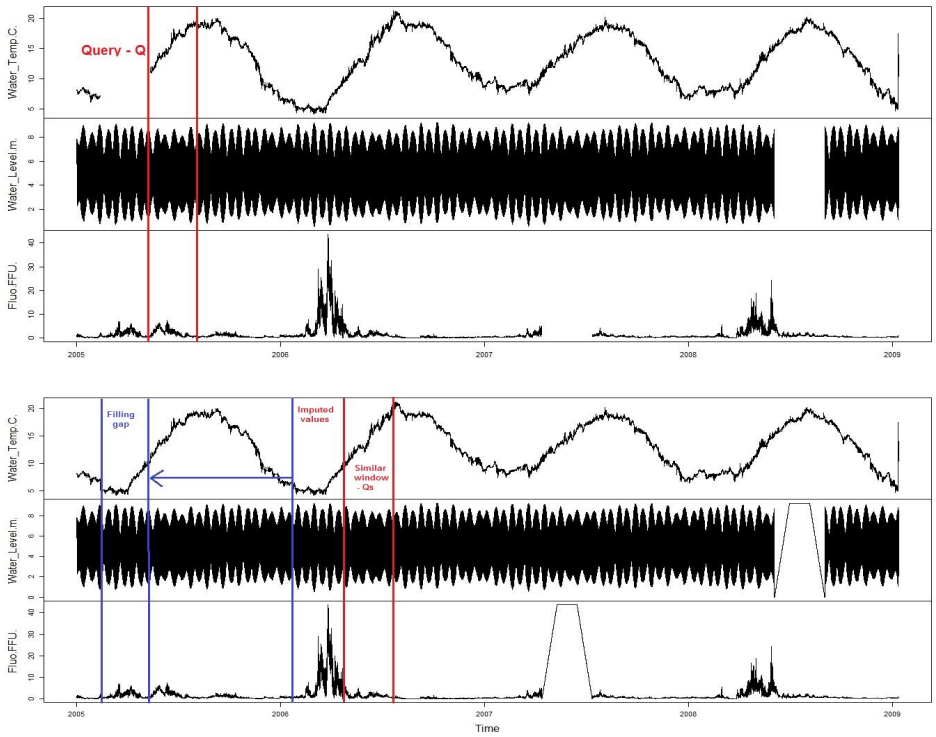


FIGURE 2.6 – Illustration du processus d'imputation du signal de température de l'eau (WaterTemp) d'une sous-série issues des données Marel-Carnot. En rouge, la construction de la requête et la recherche d'une fenêtre similaire et en bleu le segment complété

.

(complétion par moyenne, médiane ou dernière valeur observée), ce trapèze permet de garder une incertitude sur la forme du signal.

$$f(i) = \begin{cases} 0 & (i < t) \text{ or } (i > t+T) \\ (i-t)/(a-t) & t \leq i \leq a \\ 1 & (a \leq i \leq b) \\ (b-i)/(b-t-T) & (b \leq i \leq t+T) \end{cases}$$

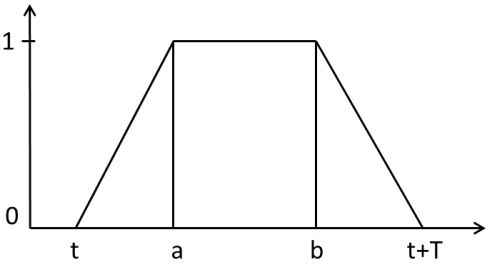


FIGURE 2.7 – Formulation de la fonction trapèze

L'approche DTWUMI a été proposé dans (T.T.H PHAN et al. 2017). Elle est particulièrement adaptée aux données dites MNAR (Missing not at Random), c'est-à-dire que les données manquantes dépendent d'autres valeurs manquantes et que les données manquantes ne peuvent pas être estimées à partir des variables existantes en un instant donné. L'estimation multivariée permet d'absorber le phénomène, la dynamique du processus et engendre une estimation acceptable. Quelques résultats seront apportés au tableau 2.7. Son inconvénient majeur reste son temps de calcul lié au calcul du coût de déformation, reporté en comparaison au tableau 2.8.

FSMUMI - Fuzzy Similarity Measure based Imputation for Univariate and Multivariate series

L'approche de comparaison basée similarité étant une voie prometteuse dans l'imputation des séries et leur classification, nous avons cherché à construire une similarité plus rapide que celle élastique et permettant pareillement d'absorber les incertitudes sur la dynamique du signal. Nous avons ainsi proposé de créer une nouvelle mesure de similarité pondérée à partir de similarités facilement calculables. Les pondérations sont estimées à partir de règles floues représentées par les figures 2.8 et 2.9.

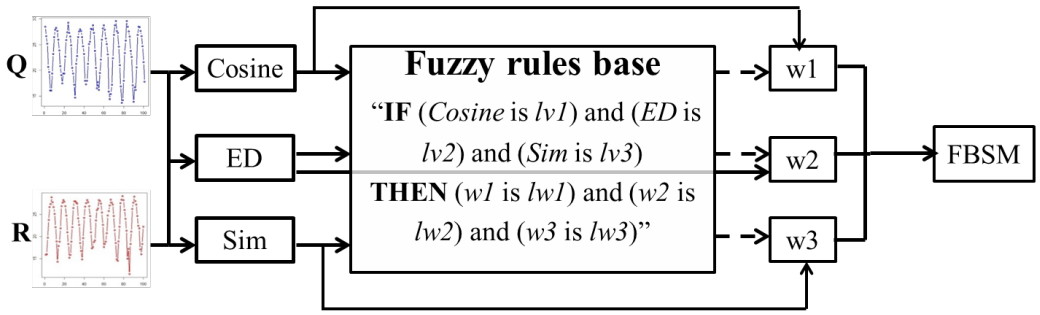


FIGURE 2.8 – Système expert calculant une similarité FBSM pondérée à partir de métriques de comparaison des signaux Q et R, et de règles floues (attention ici ED : Euclidean dissimilarity)

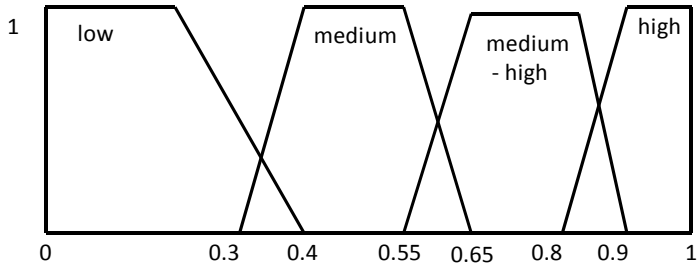


FIGURE 2.9 – Partition floue des valeurs de similarités.

Dans nos travaux (T.T.H. PHAN, BIGAND et **E. Poisson Caillault** 2018), la similarité FSBM est issue de 3 métriques pondérées et est définie par l'équation 2.12 où chaque poids sont appris $w_i, i \in 1, 2, 3$ à partir de 64 règles et $lvi, lwi \in$ {low, medium, medium-high, high} définissent les valeurs linguistiques de w_i :

$$FBSM = w_1 * Cosine(Q, R) + w_2 * ED'(Q, R) + w_3 * Sim(Q, R) \tag{2.9}$$

$$Cosine(Q, R) = \frac{\sum_{i=1}^{T} q_i . r_i}{\sum_{i=1}^{T} (q_i)^2 . \sum_{i=1}^{T} (r_i)^2} \tag{2.10}$$

$$ED^*(Q, R) = \frac{1}{1 + \sqrt{\sum_{i=1}^{T} (q_i - r_i)^2}} \tag{2.11}$$

$$Sim(Q, R) = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{1 + \frac{|q_i - r_i|}{max(Q) - min(Q)}} \tag{2.12}$$

Le système a été développé à partir du package Fuzzy-R fournissant un moteur d'inférence flou (WAGNER, MILLER et GARIBALDI 2011).

Expérimentations numériques DTWUMI et FSMUMI

Ces approches ont été comparées à plusieurs techniques d'imputation multiple (Amelia II-, MI, MICE), des estimations par classifieurs tels les forêts aléatoires ou FCM ou encore d'interpolation comme na.approx citée précédemment en imputation univarié,

— MI- Multiple Imputation (Rubin 1996 est basé sur une analyse de plusieurs tirages plausibles avec une modélisation bayésienne pour remplacer une valeur absente.

— MICE - Multivariate Imputation Chained Equations (Buuren et Groothuis-Oudshoorn 2011) est basé sur une régression logistique multiple utilisant l'algorithme Monte-Carlo Markov Chain. Il est adapté au cas de données manquantes aléatoirement (dits MA). Pour son utilisation, les valeurs manquantes sont initialement complétées par la dernière donnée observée.

— Amelia II (Honaker, King et Blackwell 2011) utilise une approche EM (expectation-maximisation). Elle est adaptée pour des séries où les valeurs manquantes dépendent des autres valeurs, dites MAR (Missing At Random), elle impose ainsi une hypothèse de distribution normale multivariée.

— missForest (D. J. Stekhoven et Buhlmann 2011) est basé sur la construction de forêts aléatoires (RandomForest) pour chaque variable. Un agent apprend la valeur de cette variable selon les valeurs insérées des autres variables. La moyenne est utilisée comme remplissage préalable. Ces agents serviront à prédire les valeurs manquantes.

— FcM - Fuzzy c-means est aussi un estimateur dit MAR basé sur classifieur flou construit sur les données réduites à celles complètes. La valeur de la variable manquante sera alors estimée à partir des centroïdes calculés et leurs degrés d'appartenances associées.

Nous avons retenu trois jeux de comparaison, l'ensemble du protocole est décrit dans le chapitre 3 de Phan 2018. Le premier jeu, appelé "Synthetic" par E.J. Keogh et M. Pazzani p.d. est dérivé de la série multivariée initialement composé de 10 signaux synthétiques de taille N=100 000. Chaque variable (signal) est une succession de structures à différentes résolutions et construite à partir de la somme de quatre sinusoides de fréquence $f_i = 2^{2+i} + rand(2^i)$ où i varie de 3 à 7. Seuls les trois premiers signaux ont été extraits pour les expérimentations (à des fins de réduction du temps de calcul). Le second sont des données simulées comprenant trois séquences de 800 observations peu corrélées issus du package-R DTWUMI (dataDTWUMI). Et, le troisième est obtenu à partir des signaux de fluorescence, niveau d'eau et température de l'eau du dataset MAREL-Carnot (Alain Lefebvre 2015), ces signaux ont été choisis pour leur corrélation faible et leur nombre de données manquantes faible.
Une synthèse des résultats est apporté dans le tableau 2.7, elle est restreinte aux indicateurs de performance pour la complétion de T-Trou de 2 % de la taille de la série. Ce choix est lié aux conclusions qui résument les comportements globaux de nos approches.

Chaque technique de complétion/imputation est appliquée aux mêmes T-Trou ; 5 T-Trou ont été créés aléatoirement. Sur les critères bornés des indicateurs, dans le cas des deux premiers jeux de données (artificiels), aucune méthode ne les remplit. Certes, la technique de complétion par FSMUMI est la plus proche (FB$<0,3$ et 1-FA2$<0,2$, FSD$<0,5$). Dans le cas du jeu de données réelles étudié, seules les approches DTWUMI et FSMUMI satisfont les limites données (1-FA2>0,2 pour les autres techniques). Les méthodes d'imputation multiple sont les moins performantes quel que soit le jeu utilisé. Certes, elles présentent des temps de calcul plus faibles (Amelia, MI) comme le montre le tableau 2.8. À noter aussi que la méthode la plus élémentaire et rapide, na.approx ne présente pas ici des résultats aberrants démontrant que ces seuls indicateurs ne suffisent pas à montrer leur efficacité à respecter les dynamiques locales des signaux notamment sur des T-Trous non monotones. Concernant les approches introduisant la prise en compte de l'incertitude explicitement soit FCM, missForest et FSMUMI, cette dernière semble la plus opportun pour les expérimentations menées tant selon une évaluation quantitative que visuelle (T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018).

La comparaison entre nos deux approches proposées dépendra fortement du critère de vitesse de complétion. Elle sera plus mitigée lorsqu'il s'agit des séries à fortes dynamiques vers une préférence pour l'approche DTWUMI avec sa vision plus globale de l'ensemble des signaux. A sa décharge, nous n'avons pas optimisé les calculs inhérents au calcul du coût DTW.

Perspectives

Bien que ces approches ne se suffisent pas elle-mêmes pour compléter une série avec des segments de valeurs manquantes et que des critères de rejet de complétion malgré une fenêtre admissible ont été paramétrés pour le moment de manière empirique, les approches par recherche d'un phénomène identique dans la série ont montré un intérêt important pour la complétion de large sous-séquence de valeurs consécutives absentes. Rappelons que la notion de "large" sous-entend que cette sous-séquence est non monotone.

Des expérimentations sur l'intégration des connaissances en amont et en aval de la séquence ont été proposées lorsque nous ne sommes pas dans un but de prédiction mais bien d'études/modélisation d'une série passée. La figure 2.10 intègre cette idée avec la construction d'une requête amont Qb (b pour before) et une requête aval Qa. Des premières propositions de fusion et de contraintes ont été soulevées et sont encore à approfondir. En effet, les fenêtres glissantes de recherche peuvent être contraintes temporellement, ce qui n'est pas le cas illustré sur ce schéma où il est toléré une recherche indépendante et supposé que la conséquence d'un phénomène ne se produit pas toujours à la même période. Dans (T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018), la manière de compléter le T-Trou est réalisée par une moyenne de la fenêtre sélectionnée suivant Qbs et celle précédent Qas. Cette fenêtre reste à définir et pourrait prendre en compte l'incertitude sur la dynamique respective des deux fenêtres ou de l'ensemble de solutions possibles telles une imputation multiple.

Un autre paramètre à considérer est la taille des fenêtres de recherche qui, par simplicité de programmation, a été fixée à la taille du T-Trou. En effet, celui-ci pourrait être variable fonction des caractéristiques fréquentielles notamment de saisonnalité du processus considéré ou des temps entre les forçages et leurs réponses.

Tableau 2.6 – Indicateurs de performance des techniques d'imputation moyennés sur 10 à 50 T-Trous réalisés, T étant fixés à 10 % de la taille de chaque série (en gras, les meilleurs scores).

	1. AirPassenger				**2. CO2 Concentration**			
Méthodes	1-Sim	NMAE	RMSE	FSD	1-Sim	NMAE	RMSE	FSD
DWTBI	**0,11**	**0,02**	**12,7**	0,36	**0,07**	**0,001**	**0,4**	**0,04**
na.interp	0,14	0,021	13,1	**0,34**	0.24	0,051	1,4	0,88
na.locf	0,21	0,042	26,1	2	0.24	0,054	1,6	2
na.prox	0,21	0,041	24,6	1,03	0,24	0,051	1,4	0,88
na.aggregate	0,19	0,035	22,1	2	0,56	0,197	4,9	2
na.spline	0,38	0,134	78,3	0,52	0,34	0,098	2,9	0,26
	2. Bearsales				**6-Mackey Glass Chaotic**			
Méthodes	1-Sim	NMAE	RMSE	FSD	1-Sim	NMAE	RMSE	FSD
DTWBI	0,16	**0,054**	1	**0,13**	**0,07**	**0,008**	**0,01**	**0,01**
na.interp	**0,11**	0,068	**0,7**	0,18	0,19	0,03	0,04	0,98
na.locf	0,18	0,13	1,3	2	0,21	0,036	0,05	2
na.prox	0,18	0,124	1,2	1,24	0,19	0,03	0,04	0,98
na.aggregate	0,16	0,111	1,1	2	0,17	0,025	0,03	2
na.spline	0,45	0,558	4,9	0,067	0,29	0,058	0,08	0,33
	3. Google				**7. Phu Lien Temperature**			
Méthodes	1-Sim	NMAE	RMSE	FSD	1-Sim	NMAE	RMSE	FSD
DTWBI	0,16	0,13	0,032	**0,23**	**0,12**	**0,063**	**1,8**	**0,05**
na.interp	0,15	0,1	0,03	1,22	0,19	0,137	3	0,58
na.locf	0,17	0,13	0,035	2	0,23	0,176	3,8	2
na.prox	0,15	0,1	0,03	1,22	0,19	0,137	3	0,58
na.aggregate	**0,13**	**0,08**	**0,024**	2	0,17	0,114	2,4	2
na.spline	0,58	4,68	1,118	1,13	0,51	0,88	17,8	1,04
	4. SP				**8. Water Level MAREL Carnot**			
Méthodes	1-Sim	NMAE	RMSE	FSD	1-Sim	NMAE	RMSE	FSD
DTWBI	0,19	0,029	40,1	0,57	**0,03**	**0,005**	**0,1**	**0,03**
na.interp	0,18	0,025	36,3	0,56	0,19	0,041	0,4	0,91
na.locf	0,19	0,026	36,9	2	0,19	0,043	0,5	2
na.prox	**0,17**	0,024	33,5	1,14	0,19	0,041	0,4	0,91
na.aggregate	0,18	**0,023**	**31,7**	2	0,17	0,036	0,4	2
na.spline	0,24	0,049	63,2	**0,45**	0,82	1,57	15,5	1,79
	9. tsNH4							
Méthodes	1-Sim	NMAE	RMSE	FSD				
DTWBI	**0,19**	**0,028**	**8,33**	**0,30**				
na.kalman	0,32	0,58	23,28	0,75				

Indicateurs	Erreur			Forme		
Synthetic	1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
FSMUMI	**0,1**	0,295	**0,046**	**0,155**	**0,395**	**0,337**
na.approx	0,104	**0,278**	0,047	0,224	0,398	0,347
FcM	0,208	0,686	0,104	1,863	2,289	0,987
DTWUMI	0,237	0,775	0,867	0,509	8,449	0,646
missForest	0,239	0,968	0,133	0,279	3,156	0,792
MICE	0,244	0,968	0,14	0,255	7,616	0,759
Amelia	0,259	0,998	0,147	0,275	2,005	0,803
MI	0,259	0,998	0,147	0,268	2,11	0,81
Simulated	1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
FSMUMI	**0,068**	**0,487**	**1,166**	0,194	1,971	0,611
DTWUMI	0,074	0,523	1,545	**0,008**	3,686	**0,583**
FcM	0,093	0,999	1,672	1,985	**1,96**	0,998
missForest	0,096	1	1,769	0,941	2,777	0,858
na.approx	0,118	1	2,261	0,721	2,059	0,786
MICE	0,119	0,999	2,282	0,114	8,881	0,789
Amelia	0,120	0,998	2,312	0,107	2,191	0,794
MI	0,120	1	2,307	0,123	3,949	0,789
MAREL-Carnot	1-Sim	1-R^2	RMSE	FSD	FB	1-FA2
DTWUMI	**0,042**	**0,018**	**1,095**	**0,029**	**0,066**	**0,154**
FSMUMI	0,045	0,037	1,446	0,053	0,083	0,182
na.approx	0,06	0,07	2,012	0,045	0,094	0,214
FcM	0,116	0,06	3,418	0,415	0,237	0,231
missForest	0,116	0,155	3,575	0,33	0,193	0,258
MICE	0,129	0,369	4,711	0,197	0,21	0,413
MI	0,146	0,364	4,72	0,218	0,228	0,435
Amelia	0,146	0,369	4,743	0,211	0,222	0,429

Tableau 2.7 – Indicateurs de qualité des techniques d'imputation sur des critères d'écart et de forme. Résultats moyennés sur un tirage de 5 T=2%-Trou de positions aléatoires. En gras, les résultats optimaux sont mis en évidence.

	Taille relative du T-Trou pour N=100 000 points						
Méthode	1 %	2 %	3 %	4 %	5 %	7,5 %	10 %
na.approx	**0,1**	**0,1**	**0,2**	**0,1**	**0,1**	**0,1**	**0,1**
Amelia	3,2	3,4	5,2	3,2	3,2	3,2	3,2
FcM	40,9	39,8	40,0	41,1	41,2	46,7	45,6
MI	844,1	714,0	739,1	723,3	724,5	719,7	726,5
FSMUMI	353,9	427,5	701,9	1 037,8	1 423,6	2 525,5	3 556,8
MICE	7 021,1	9 187,7	21 909,6	13 041,9	14 833,9	19 417,7	23 812,6
missForest	26 833,8	24 143,8	22 969,9	32 056,6	36 485,8	42 424,1	28 521,1
DTWUMI	5 002,7	15 714,8	37 645,8	64 669,7	86 435,4	180 887,8	273 879,0

Tableau 2.8 – Temps de calcul moyen en seconde (s) pour compléter des T-trou dans la série "synthetic" selon la taille T et l'algorithme utilisé.

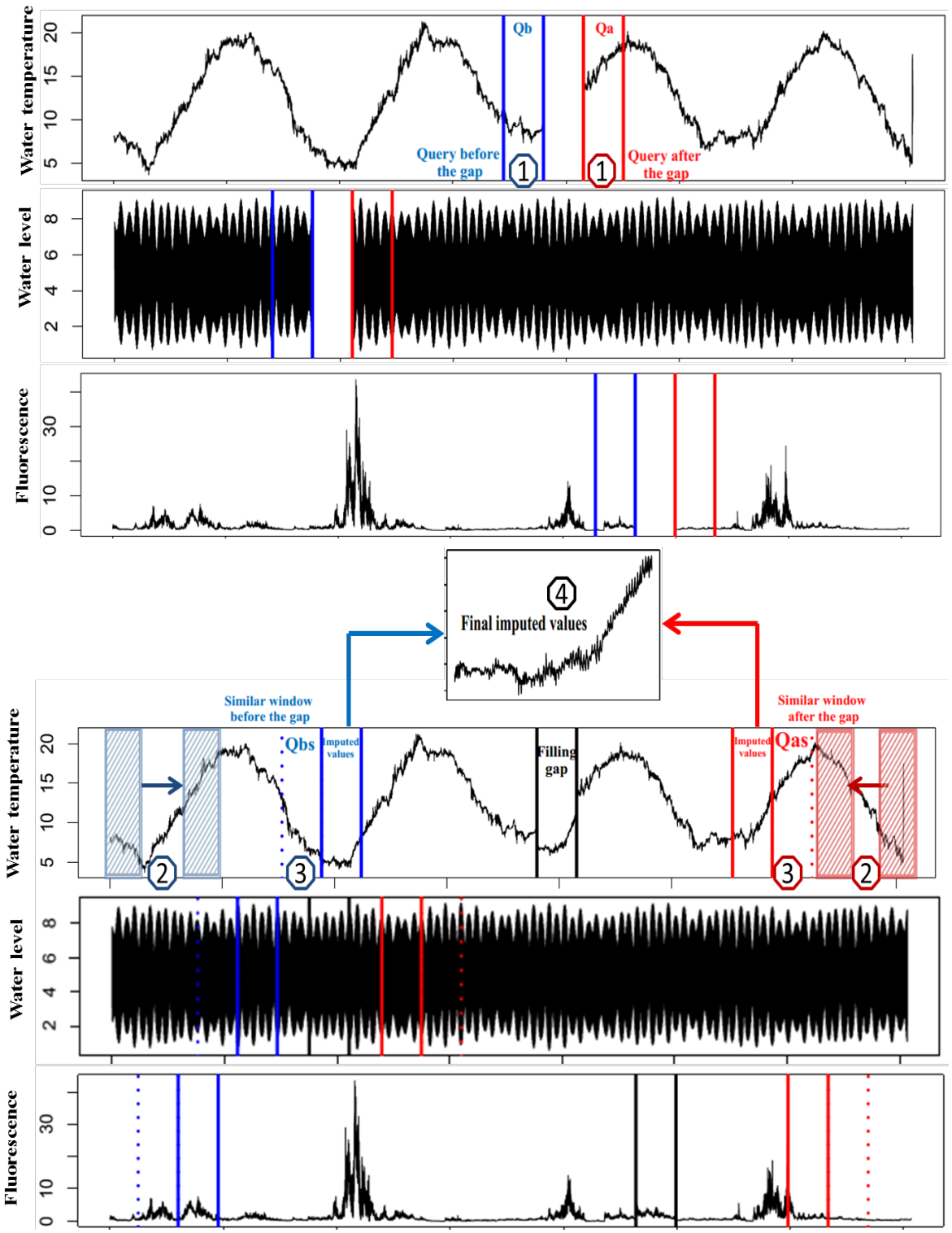


FIGURE 2.10 – Complétion par recherche d'un phénomène passé et futur au T-Trou : 1-Construction des requêtes ; 2- Balayage des fenêtres similaires ; 3- Sélection de la ou les fenêtres admissibles ; 4 - fusion de ces fenêtres pour compléter le T-trou.

2.3 Classification et Partitionnement de séries

Dans cette partie, nous nous intéressons maintenant au volet classification. L'angle d'attaque choisi est toujours basé sur la similarité entre celles-ci et plus précisément sur l'utilisation du spectre de cette similarité et la minimisation d'un critère de coupe (spectral clustering).

2.3.1 Un point de vue sur les techniques de classification spectrale

Les approches par classification spectrale permettent de lever des hypothèses sur la géométrie et distribution des données.

La figure 2.11 tirée de la page de Bob Baxley (http://rjbaxley.com - site très illustré sur les approches par graphe et leurs applications) illustre cette affirmation. En effet, les approches conventionnelles de représentations basées sur des centres et des densités de points, tels les modèles de mélanges gaussiens (GMM, k-means), imposent des nuages de points appartenant au même cluster convexe et clairement séparable. Les approches hiérarchiques divisives (Birch, Ward) requièrent soit de connaître le nombre K de groupe souhaité, soit de descendre fortement dans la hiérarchie puis de chercher des critères d'agglomérations de ceux-ci pour obtenir des clusters cohérents visuellement. Ils sont très sensibles lorsque les clusters sont connexes et ne produisent pas toujours une coupe cohérente comme il est illustré sur l'exemple des trois nuages/traits obliques à la 4ème ligne de la figure 2.11. Les approches agglomératives dont DBSCAN (Density-Based Spatial Clustering of Applications with Noise) requièrent un paramétrage de l'écart admissible entre deux points d'un même cluster. Elles sont intéressantes puisqu'elles ne nécessitent pas la connaissance de K et ne forceront pas l'appartenance à un cluster : certains points seront considérés non classables, dit plus communément outliers. L'algorithme des k-moyennes (k-means Hartigan et Wong 1979) ou encore des k-médoïdes ne permettent pas de séparer des clusters convexes, pour outrepasser ce verrou, plusieurs classifieurs projettent initialement les données dans un espace "idéal" avant de rechercher ces centres représentants pour former une partition. Tel est le cas ici de Mean-shift basé sur un espace noyau, la classification spectrale basée sur les vecteurs propres (spectre) de la matrice de similarité ou encore la méthode AP (Affinity Propagation) basée sur une similarité de voisinage.

Bien que nécessitant quelques paramétrages (choix du nombre K de cluster, du critère de coupe, du type de similarité), les approches par clustering spectral (SC - Spectral Clustering) sont très efficaces dans grands nombres de domaines (Suryanarayana, Rao et Swamy 2015 ; Filippone et al. 2008). Elles ont eu un regain d'intérêt tant sur le paramétrage que sur l'accélération du temps de calcul (Yan, Huang et Jordan 2009 ; Langone et Suykens 2017).

En classification spectrale, la géométrie des données se traduit par un graphe non orienté noté $g = (V, E, W)$ où $V = \{x_1, ..., x_n\}$ est l'ensemble des noeuds correspondant aux points x_i ou séries à classer,E les liaisons entre ces points, W la force de cette liaison c'est-à-dire leurs similarités. Typiquement, ces similarités W_{ij} sont des matrices de voisinage (W_{ij}=1 si le point x_j appartient au voisinage du point x_i, zéro sinon) ou définies à partir d'un noyau gaussien selon l'équation $W_{ij} = e^{-\frac{||x_i - x_j||_2}{2\sigma^2}}$ où σ est un paramètre de dispersion à définir qui permettra de creuser la matrice et faciliter la séparation des données. Il peut aussi être estimé localement en utilisant la notion de voisinage ainsi $\sigma = \sigma_i \times \sigma_j$ avec σ_i la distance du point x_i à son n-ième voisin (Zelnik-Manor et Perona 2005).

Une fois le graphe défini G(V,E,W) et le critère de coupe choisi, l'optimisation de celui-ci est obtenue en résolvant un problème de valeurs propres généralisé $Lz = \lambda z$ avec L le laplacien et (λ,z) le spectre cette matrice. Nous renvoyons le lecteur au tutoriel de Von Luxburg (Luxburg 2007) ou encore au chapitre 2 de la thèse de Guillaume Wacquet 2011 qui décrit l'ensemble des critères et des Laplaciens associés, et reprend les démonstrations de passage du critère de coupe à sa résolution. Nous résumons le processus K-partition par

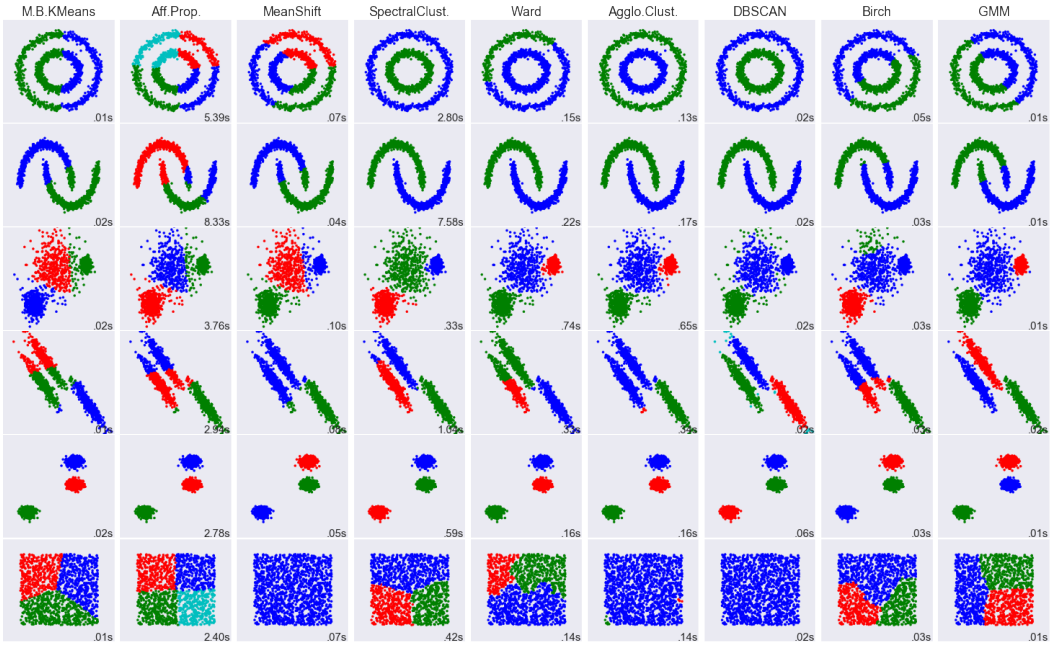


FIGURE 2.11 – Comparaison des techniques de clustering sur plusieurs jeux de données générés tiré de Bob Baxley http://rjbaxley.com. (MB Kmeans : Minibatch Kmeans ; Aff. Prop. : Affinity Propagation, Agglo. Clust : Agglomerative Clustering, DBSCAN : Density Based Spatial Clustering of Applications with Noise, GMM : Gaussian Mixture Model)

classification spectrale dans l'algorithme 3 en introduisant l'approche proposée par NJW (NG, JORDAN et WEISS 2002) et un critère multi-coupe normalisé $J_{SC} = \sum_{(i,j)}(u_i - u_j)^2 W_{ij}$ relaché par $J_{SC} = z^T L z$ avec $u = D^{-1/2} z$, u étant un vecteur indicateur d'appartenance au cluster.

Algorithme 3 Algorithme K-partition par classification spectrale (SC - Spectral clustering).

Require: W(X) : une matrice de similarité Gram des points, K : nombre de groupe.

1: (optionnel) $W_{ij} = 0$
2: Construire D la matrice des degrés de W.
3: Calculer son laplacien L : $L = D^{-1/2} W D^{-1/2}$.
4: Extraire le spectre de L (λ,z).
5: Former $Z = \{z_1, \ldots, z_K\}$ les K plus grands vecteurs propres au sens de Λ.
6: (optionnel) Normaliser les lignes de Z par projection sur une sphère unité $F_{ij} = Z_{ij} / \sum_j Z_{ij}^2$.
7: Appliquer un algorithme de partitionnement sur les lignes de F (ou F=Z si non).
NJW : K-means
8: Assigner pour chaque objet X_i le groupe de F_i obtenu dans l'espace spectral.
9: **return** vecteur de labels Y=label(X).

Pour $K = 2$, l'opérateur signe est utilisé pour identifier le label des points. Avec $K \geq 2$, l'algorithme K-means ou sa version robuste K-médoides sont des plus utilisées. Le mot spectral-Kmeans est alors souvent employé pour désigner la technique de classification spectrale.

2.3.2 Intégration de connaissances dans le processus de clustering.

La classification peut être guidée par des connaissances *a priori*. Ces connaissances sont de deux types : étiquetées faisant intervenir un expert ou géométriques. Nous qualifierons d'expert une personne ayant la compétence d'affirmer que la série ou le point est de telle type

ou une machine ayant acquis cette compétence par apprentissage.

Lorsqu'une partie seulement des labels des données est connue, nous sommes dans un cadre dit semi-supervisé. Deux approches se distinguent dans la littérature : la propagation des labels dans le vecteur des labels estimés ou par transformation de la matrice de similarité. Nous résumons sommairement leurs philosophies et renvoyons au livre de Chapelle, Schlkopf et Zien 2010. Dans les algorithmes dit de propagation des labels, le vecteur des labels estimés de l'ensemble des points Y est introduit directement avec $\hat{Y} = (\hat{Y}_l, \hat{Y}_u)$ où $\hat{Y}_l$ représentent la partie des labels connus initialement et $\hat{Y}_l$ les données non étiquetées. Ce vecteur est initialisé avec $Y^{(0)} = (y_1, \ldots, y_l, 0, \ldots, 0)$ puis recalculé jusqu'à convergence selon le critère visé avec une pondération $\alpha \in [0, 1]$ entre le poids de la géométrie et les labels connus introduite dans le laplacien L : $\hat{Y}^{t+1} = \alpha L \hat{Y}^t + (1 - \alpha) Y^{(0)}$. A noter que selon les approches, $\hat{Y}_l$ peut différer des labels initiaux fournis Y_l. La seconde approche consiste à modifier la matrice de similarité en forçant la partie W_{ll}. Diverses techniques utilisent alors des régularisation de graphes ou de marches aléatoires (Markov Random Walks) dans un graphe avec des transitions imposées par les labels.

Cette idée de forçage de la matrice W peut être repris aisément dans l'introduction de contraintes géométriques. $W = f(W^{(0)}, Wg)$ où $W^{(0)})$ est la matrice de similarité classique et Wg la matrice tenant compte des contraintes géométriques imposées par un graphe de ϵ-voisinage, (mutuel-) *kppv*-graphe, de corrélation (Lucińska et Wierzchoń 2012 ; Amizadeh 2014) ou des graphes obtenus par des techniques de réduction et préservation des distances (LPP - Locally preserving Projections, LE - Laplacian Eigenmap, ..., Fu et Ma 2013).

Une autre approche est de considérer des approches par paire introduites par Wagstaff, Davidson et Basu 2008. La connaissance acquise étant parfois limitée, l'expert ou tout autre personne par visualisation pourra apporter des informations sommaires et non exactes (comme une étiquette spécifique) telles deux séries ou deux données se ressemblent ou ne se ressemblent. Dans ce cas, un formalisme "ML=Must-Link/CNL=Cannot Not Link" (doit être liés/ne doivent pas être liés) est adopté et la matrice de similarité devient :

$$W'_{ij} = \begin{cases} +1 & \text{si } (x_i, x_j) \in C_{ML} ie. \text{doivent appartenir au même cluster} \\ 0 & \text{si } (x_i, x_j) \in C_{CNL} ie. \text{ne doivent pas appartenir au même cluster} \\ W_{ij} & \text{sinon.} \end{cases}$$

Ce formalisme a été inséré dans un partitionnement K-means (Wagstaff, Cardie et al. 2001) et donc facilement transposable dans un cadre spectral-K-means. Kamvar, Klein et Manning 2003 est le premier à l'avoir retenu explicitement, la matrice W' se substitue à la matrice W en entrée de l'algorithme SC. Wang et Davidson 2010 ont proposé d'insérer les contraintes dans le critère d'optimisation dans le critère de coupe avec une matrice de contraintes $Q = \{q_{ij}\}$ défini par :

$$q_{ij} = \begin{cases} +1 & \text{si } (x_i, x_j) \in C_{ML} \\ -1 & \text{si } (x_i, x_j) \in C_{CNL} \\ 0 & \text{sinon.} \end{cases}$$

Le critère de coupe $J_{FCSC} = J_{SC}$ sous contrainte que $u^T Q u = \sum_{i,j} u_i u_j q_{ij}$

2.3.3 Critère multi-coupe normalisé intégrant des contraintes de comparaison par paires.

Dans les travaux de Guillaume Wacquet 2011, nous avons aussi revisité cette idée d'un matrice de contrainte et adapté l'algorithme de clustering spectral NJW pour introduire des connaissances *a priori* sous la forme d'appartenance (ou non) aux mêmes clusters. Le critère

multicoupe normalisé nommé J_{cSC} (Constrained Spectral Clustering - classification spectrale contrainte (cSC)) introduit une pondération qui contrebalance le poids entre le respect des contraintes et la géométrie des données. Ce critère est alors défini par :

$$J_{cSC} = \gamma \sum_{(i,j)} (u_i - u_j)^2 W_{ij} + (1-\gamma) \sum_{(i,j)} (u_i - u_j)^2 Q_{ij} \tag{2.13}$$

$$= \sum_{(i,j)} (u_i - u_j)^2 (\gamma W_{ij} + (1-\gamma) Q_{ij}) \tag{2.14}$$

$$= \sum_{(i,j)} (u_i - u_j)^2 W_{cSC} \tag{2.15}$$

$$= z^T L_{cSC} z \text{ avec } z = D^{-1/2} u \tag{2.16}$$

La matrice $W_{cSC} = (\gamma W_{ij} + (1-\gamma) Q_{ij})$ est donc introduite directement en entrée de l'algorithme SC pour fournir une partition des données. Le paramètre γ permet de soulager le respect des contraintes qui pourraient contredire la structure du nuage de points et aboutir à aucun partitionnement possible.

Expérimentations numériques

Dans G. Wacquet, **E. Caillault Poisson** et al. 2013, le formalisme cSC a été comparé à diverses approches de clustering constraint : cPCA - constrained Principal Component Analysis, cLPP constrained Locality Projection ou cLE- constrained Laplacien Eigenmap. Ses performances ont été comparées sur plusieurs bases UCI avec deux critères 'agreement', 'accurary' en fonction du pourcentage de contraintes insérées ('% of Constraints'). 'accuracy' est le taux moyen de reconnaissance des labels (puisque les labels de ces bases sont connus) et 'agreement' est une mesure de concordance de paires entre deux partitions. Il permet de comptabiliser à la fois les paires associées au même cluster dans chaque partition et celles non associées pour les deux partitions sans avoir besoin de connaître les labels de la paire de points. Nous reprenons ici uniquement ce dernier critère non supervisé nommé aussi Rand index (Rand 1971).

La figure 2.12 reprend les résultats pour 6 bases connues. Le nombre de groupes K ici a été fixé par le nombre de labels. La pondération entre la géométrie et le respect des contraintes est choisie sans *a priori* (équirépartie $\gamma = 0.5$). La matrice de similarité est calculée à partir du noyau gaussien où σ est ajusté localement aux produits des distances des septièmes voisins. Pour les algorithmes basés sur une réduction des données (cPCA,cLPP), la variance totale est fixée à 95 %.

L'approche cSC permet d'approcher rapidement la partition idéale donnée par celles de la vérité terrain (labels donnés dans UCI Dua et Graff 2017) avec peu de connaissance par paire apportée. Dans (G. Wacquet, **E. Poisson-Caillault** et Hebert 2013), une détermination automatique du nombre K a été proposée et étudiée fonction des contraintes et du gap entre les valeurs propres.

Application au partitionnement de séries

Dans (Gogolou et al. 2018), il a été montré que l'approche de comparaison par paires de séries dépend fortement de la représentation choisie. Notamment une comparaison entre des séries initiales peut conduire à une interprétation différente d'une comparaison entre séries déformées par méthodes élastiques si ces patrons sont similaires ou non.
Nous nous sommes intéressés, dans le cadre du projet INTERREG IVa 2 mers DYMAPHY (2010-2014), au partitionnement de séries issues d'un cytomètre en flux. Celui-ci capture des mesures de lumière réémise par diffusion et de fluorescence pour chaque cellule détectée. La vitesse du flux liquide passant devant un faisceau laser est supposée très élevée et garantir que chaque mesure caractérise une seule particule. En sortie, pour chaque cellule

détectée dans l'échantillon d'eau mesuré, une série multivariée appelée cytogramme est enregistrée. La figure 2.13 reprend des exemples de cytogrammes obtenus pour quatre espèces différents. L'objectif de nos travaux a été de proposer un outil permettant de partitionner ces cytogrammes issus d'un ou plusieurs échantillons traités en routine et de visualiser les caractéristiques des groupes obtenus. Ce partitionnement est ensuite soumis à un expert dans une tache d'étiquetage et d'amélioration en insérant des contraintes ML/CNL par visualisation de quelques cytogrammes proposés aux frontières des clusters. Cette classification guidée permet de traiter un grand nombre d'échantillons à des fréquences inférieures au temps entre deux acquisitions (10 minutes). Par ailleurs, elle apporte rapidement une solution et un découpage cohérent pour des clusters connexes et parfois peu denses, par rapport à un clustering manuel réalisés à la souris à partir de quelques vues bi-attributs, demandant un temps bien supérieur et trop fortement lié aux critères de densités des groupes obtenus.

Une première étude a été menée pour déterminer la représentation à considérer (EANN'2009 : **E. Caillault**, Hebert et Wacquet 2009 , STIC'2009 : **Caillault** et al. 2009). Une approche originale a été proposée : apprendre à classer les particules non pas à partir de leur cytogramme ou attributs dérivés mais à partir des similarités entre leur cytogramme et un ensemble d'empreintes types. Ces similarités ont été basées sur des coûts d'appariements multi-conjoints des séries incluses dans le cytogramme.
A partir de ces similarités élastiques dans (G. Wacquet, Hebert et al. 2011), une approche semi-supervisée a été conduite avec la propagation des labels par une approche par contrainte cSC. Les expérimentations montrent une supériorité de cet algorithme vis-à-vis de ceux étudiés (et du moment) à la fois en terme quantitatif (indice de Rand, F-Score, silhouette) mais aussi qualitatif avec une visualisation plane apportée à l'expert permettant de simplifier sa tâche de labellisation par paire ou étiquette. Ceci est d'ailleurs proposé dans un paquet R nommé RClusTool (`http://mawenzi.univ-littoral.fr/RclusTool/`) et testé dans les ateliers du réseau RESOMAR et du projet H2020 JERICO-Next (**Poisson Caillault** et Hébert 2013 ; A. Lefebvre et al. 2016 ; Artigas et al. 2015 ; Wacquet et al. 2018).

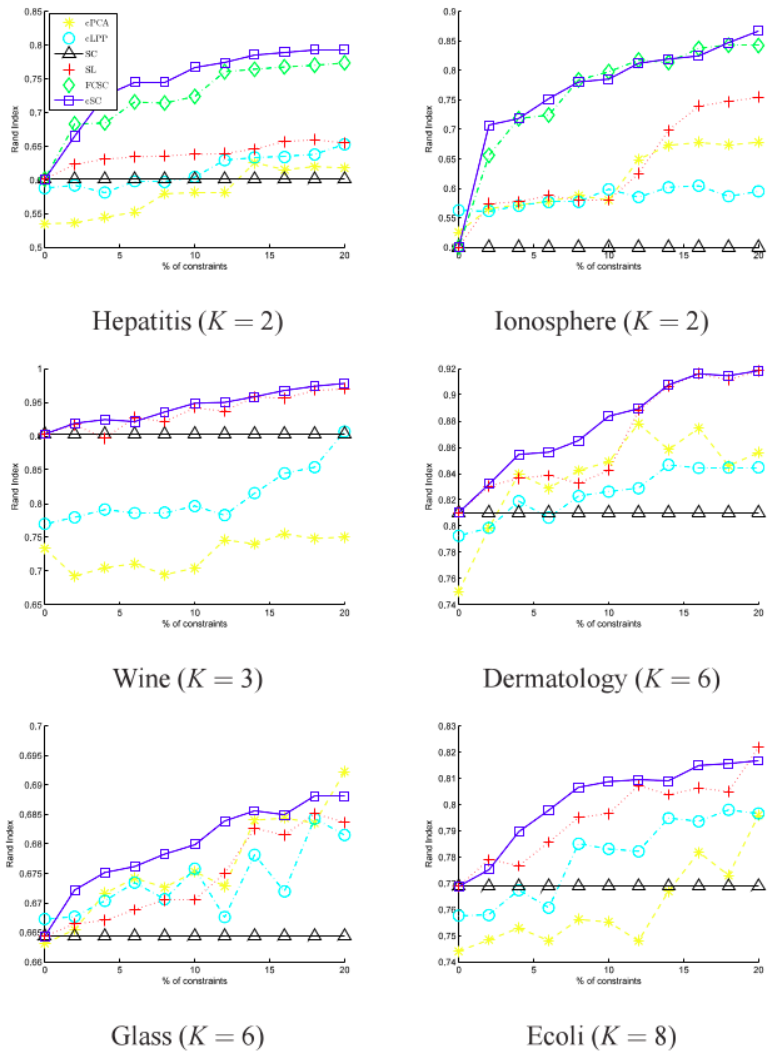


FIGURE 2.12 – Mesures de Rand Index sur différentes bases UCI (DUA et GRAFF 2017) et différents algorithmes de clustering non constraints (SC :spectral clustering) et contraints (SL basé kmeans contraint, cSC : constrained SC, FCSC Flexible Constrained SC, cPCA réduction par composantes principales contraintes, cLPP réduction contrainte par projection préservant les proximités locales entre points).

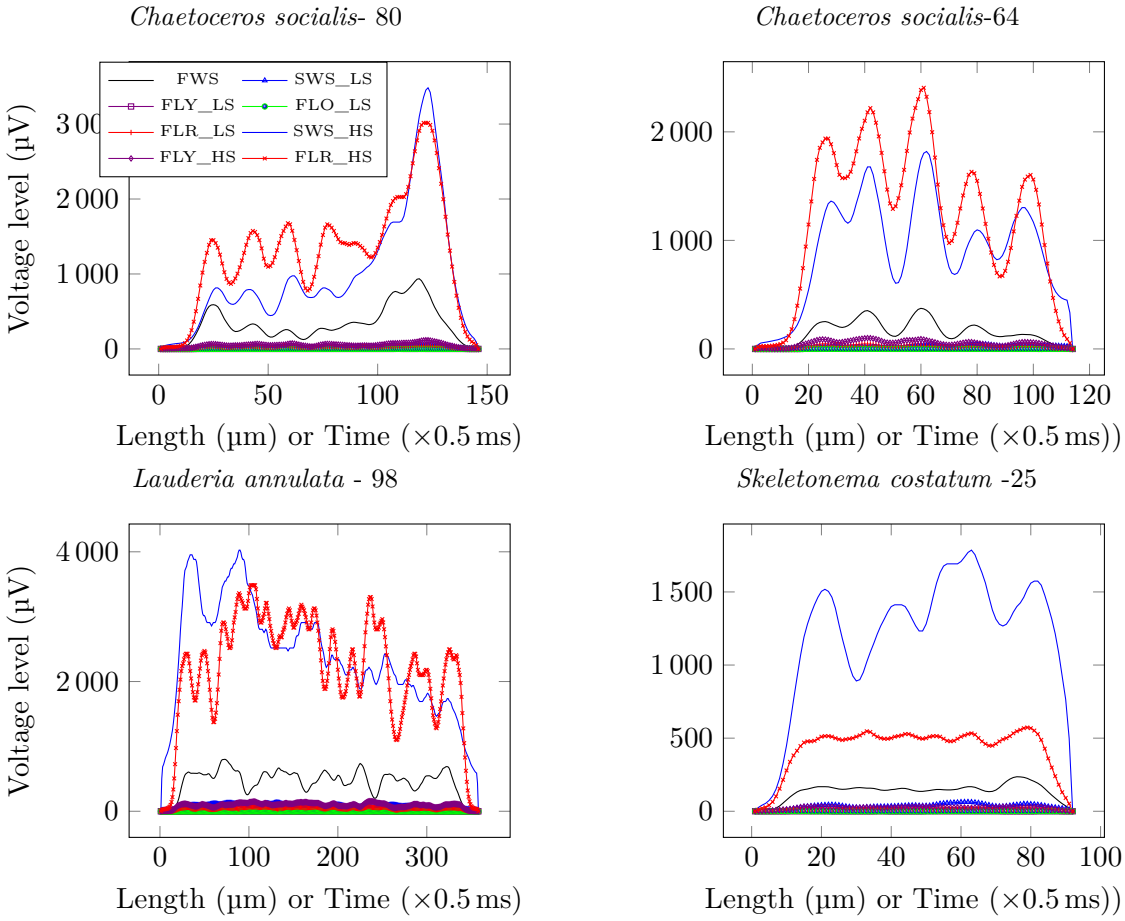


FIGURE 2.13 – Cytogramme 8D de 3 espèces phytoplanctoniques différentes.

2.4 Conclusions, Perspectives.

L'insertion de connaissances est toujours un domaine très actif, les technologies et l'internet des objets (iOT) évoluant dans ce sens offrant aujourd'hui une résolution de plus en plus fine (spatiale, temporelle ou taxonomique) et une capacité mémoire ou calcul plus importante. De la sélection des attributs et des espaces de représentations, aux perfectionnement et combinaison d'algorithmes de clustering, divers papiers fleurissent régulièrement dans la littérature.
La fusion des représentations et la visualisation intelligente de celles-ci dans le processus d'incorporation de connaissances expertes dans la phase de comparaison de données est un sujet qui reste à creuser. Les approches floues ont été introduites sommairement dans les travaux (T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018) et sont aussi à développer afin de bien mesurer l'impact de ces représentations et corriger celles-ci dans les processus visés tels la classification non supervisé, le clustering contraint ou la prédiction .
Les approches ou métriques élastiques ne sont plus à contourner aujourd'hui, de nombreux algorithmes s'attachant à accélérer leur temps de calcul et démontrer leur efficacité.

Je profite pour conclure ce chapitre de remercier Keogh et ses collaborateurs proches pour leurs "survey" et critiques constructives -http://www.cs.ucr.edu/%7Eeamonn/LB_Keogh.htm,http://www.cs.ucr.edu/~eamonn/) et leur façon d'appréhender des résultats numériques. Je reprendrai donc ce passage tiré des diapositives d'introduction des bases UCR (Chen, Eamonn Keogh et al. 2015) :

" Several researchers have published papers on showing “we win some, we lose some” on the UCR Archive. However, there are many trivial ways to get “win some, lose some” ... Gustavo Batista has pointed out that “win some, lose some” is worthless unless you know in advance which ones you will win on ! ...
It could be argued that the goal of researchers should be to solve real world problems, and that improving accuracy on the UCR Archive is at best a poor proxy for such real world problems."

Bibliographie

Allison 2001

Allison, Paul D. (2001). *Missing Data*. T. 136. Quantitative Applications in the Social Sciences. Sage Publication, p. 104.

Amizadeh 2014

Amizadeh, Saeed (2014). « Non-parametric graph-based methods for large scale problems ». Thèse de doct. University of Pittsburgh.

Artigas et al. 2015

Artigas, L. F. et al. (2015). *On the combination of semi-automated approaches and tools for measuring phytoplankton dynamics in coastal waters: implications for monitoring networks*. ASLO 2015 Aquatic Sciences Meeting. Granada, Spain.

Bishop 2006

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc. isbn : 0387310738.

Buuren et Groothuis-Oudshoorn 2011

Buuren, Stef van et Karin Groothuis-Oudshoorn (2011). « mice: Multivariate Imputation by Chained Equations inR ». In : *Journal of Statistical Software* 45.3. doi : 10.18637/jss.v045.i03. url : https://doi.org/10.18637/jss.v045.i03.

Caillault et al. 2009

Caillault, E. et al. (2009). « Classification de cytogrammes par appariement élastique. Vers la discrimination automatique du Phytoplancton marin par cytométrie en flux ». In : *Actes du colloque STIC et Environnement*. Calais, France, p. 1–13.

Ceong, Kim et Park 2012

Ceong, Hee-Taek, Hae-Jin Kim et Jeong-Seon Park (2012). « Discovery of and Recovery from Failure in a Costal Marine USN Service ». In : *Journal of information and communication convergence engineering* 10.1, p. 11–20.

Chapelle, Schlkopf et Zien 2010

Chapelle, Olivier, Bernhard Schlkopf et Alexander Zien (2010). *Semi-Supervised Learning*. 1st. The MIT Press.

Chen, Hu et al. 2013

Chen, Yanping, Bing Hu et al. (2013). « DTW-D: Time Series Semi-supervised Learning from a Single Example ». In : *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA : ACM, p. 383–391. isbn : 978-1-4503-2174-7. doi : 10.1145/2487575.2487633. url : http://doi.acm.org/10.1145/2487575.2487633.

CHEN, Eamonn KEOGH et al. 2015
CHEN, Yanping, Eamonn KEOGH et al. (2015). *The UCR Time Series Classification Archive. URL www.cs.ucr.edu/ eamonn/time_series_data/.*

DENG et al. 2016
DENG, Yi et al. (2016). « Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data ». In : *Scientific Reports* 6, p. 21689. ISSN : 2045-2322. (Visité le 19/09/2016).

DUA et GRAFF 2017
DUA, Dheeru et Casey GRAFF (2017). *UCI Machine Learning Repository.* URL : http://archive.ics.uci.edu/ml.

E. Caillault, HEBERT et WACQUET 2009
E. Caillault, PA. HEBERT et G. WACQUET (2009). « Dissimilarity-Based Classification of Multidimensional Signals by Conjoint Elastic Matching: Application to Phytoplanktonic Species Recognition ». In : *Engineering Applications of Neural Networks - 11th International Conference, EANN 2009, London, UK, August 27-29, 2009. Proceedings*, p. 153–164. DOI : 10.1007/978-3-642-03969-0_15.

FILIPPONE et al. 2008
FILIPPONE, Maurizio et al. (2008). « A survey of kernel and spectral methods for clustering ». In : *Pattern Recognition* 41.1, p. 176–190. DOI : 10.1016/j.patcog.2007.05.018. URL : https://doi.org/10.1016/j.patcog.2007.05.018.

FU et MA 2013
FU, Yun et Yunqian MA, éds. (2013). *Graph Embedding for Pattern Analysis.* Springer New York. DOI : 10.1007/978-1-4614-4457-2. URL : https://doi.org/10.1007/978-1-4614-4457-2.

G. WACQUET, **E. Caillault Poisson** et al. 2013
G. WACQUET, **E. Caillault Poisson** et al. (2013). « Constrained spectral embedding for K-way data clustering ». In : *Pattern Recognition Letters* 34.9. Impact Factor:1.062, ERA2010: B, h5=, p. 1009–1017. DOI : 10.1016/j.patrec.2013.02.003.

G. WACQUET, **E. Poisson-Caillault** et HEBERT 2013
G. WACQUET, **E. Poisson-Caillault** et PA. HEBERT (2013). « Semi-supervised K-Way Spectral Clustering with Determination of Number of Clusters ». In : *Computational Intelligence: Revised and Selected Papers of the International Joint Conference, IJCCI 2011, Paris, France, October 24-26, 2011.* Sous la dir. de Kurosh MADANI et al. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 317–332. ISBN : 978-3-642-35638-4. DOI : 10.1007/978-3-642-35638-4_21.

G. WACQUET, HEBERT et al. 2011
G. WACQUET, PA. HEBERT et al. (2011). « Classification semi-supervisee pour l'identification de cellules phytoplanctoniques ». In : *STIC et Environnement, Colloque Sciences et Techniques de l'Information et de la Communication Pour l'Environnement.*

GELMAN et al. 2015
GELMAN, Andrew et al. (2015). *Mi: Missing Data Imputation and Model Checking.* (Visité le 19/09/2016).

GenTXWarper p.d.
GenTXWarper (p.d.). *Dynamic Time Warping algorithm for gene expression time series. Available: http://www.psb.ugent.be/cbd/papers/gentxwarper/DTWalgorithn*

Gogolou et al. 2018
Gogolou, A. et al. (2018). « Comparing Similarity Perception in Time Series Visualizations ». In : *IEEE Transactions on Visualization and Computer Graphics*, p. 1–1. ISSN : 1077-2626. DOI : 10.1109/TVCG.2018.2865077.

Guillaume Wacquet 2011
Guillaume Wacquet (2011). « Classification spectrale semi-supervisée : Application à la supervision de l'écosystème marin. (Constrained spectral clustering: Application to the monitoring of the marine ecosystem) ». Thèse de doct. supervised by D. Hamad et E. Caillault, University du Littoral, Dunkerque, France.

Hartigan et Wong 1979
Hartigan, J. A. et M. A. Wong (1979). « Algorithm AS 136: A K-Means Clustering Algorithm ». In : *Applied Statistics* 28.1, p. 100. DOI : 10.2307/2346830. URL : https://doi.org/10.2307/2346830.

Honaker, King et Blackwell 2011
Honaker, James, Gary King et Matthew Blackwell (2011). « AmeliaII: A Program for Missing Data ». In : *Journal of Statistical Software* 45.7. DOI : 10.18637/jss.v045.i07. URL : https://doi.org/10.18637/jss.v045.i07.

Hyndman et Khandakar 2008
Hyndman, Rob J. et Yeasmin Khandakar (2008). « Automatic Time Series Forecasting: TheforecastPackage forR ». In : *Journal of Statistical Software* 27.3. DOI : 10.18637/jss.v027.i03. URL : https://doi.org/10.18637/jss.v027.i03.

Joseph et al. 2009
Joseph, Jill G. et al. (2009). « Reducing Psychosocial and Behavioral Pregnancy Risk Factors: Results of a Randomized Clinical Trial Among High-Risk Pregnant African American Women ». In : *American Journal of Public Health* 99.6, p. 1053–1061. ISSN : 0090-0036. (Visité le 20/09/2016).

Junninen et al. 2004
Junninen, Heikki et al. (2004). « Methods for Imputation of Missing Values in Air Quality Data Sets ». en. In : *Atmospheric Environment* 38.18, p. 2895–2907. ISSN : 13522310. (Visité le 19/09/2016).

Kamvar, Klein et Manning 2003
Kamvar, Sepandar D., Dan Klein et Christopher D. Manning (2003). « Spectral learning ». In : *In IJCAI*, p. 561–566.

E. J. Keogh et M. J. Pazzani 2001
Keogh, Eamonn J. et Michael J. Pazzani (2001). « Derivative Dynamic Time Warping ». In : *In First SIAM International Conference on Data Mining (SDM'2001)*.

E.J. Keogh et M. Pazzani p.d.
Keogh, E.J. et M.J. Pazzani (p.d.). « An indexing scheme for fast similarity search in large time series databases ». In : *Proceedings. Eleventh International Conference on Scientific and Statistical Database Management*. IEEE Comput. Soc. doi : 10.1109/ssdm.1999.787621.

Kévin Rousseeuw 2014
Kévin Rousseeuw (2014). « Modélisation de signaux temporels hautes fréquences multicapteurs à valeurs manquantes : Application à la prédiction des efflorescences phytoplanctoniques dans les rivières et les écosystèmes marins côtiers ». Thèse de doct. supervised by E. Caillault et A. Lebfebvre, dir. D. Hamad, Université du Littoral Côte d'Opale, Boulogne-sur-Mer, France.

Langone et Suykens 2017
Langone, Rocco et Johan A.K. Suykens (2017). « Fast kernel spectral clustering ». In : *Neurocomputing* 268. Advances in artificial neural networks, machine learning and computational intelligence, p. 27–33. issn : 0925-2312. doi : https://doi.org/10.1016/j.neucom.2016.12.085. url : http://www.sciencedirect.com/science/article/pii/S0925231217307488.

Lee et Carlin 2010
Lee, Katherine J. et John B. Carlin (2010). « Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation ». en. In : *American Journal of Epidemiology* 171.5, p. 624–632. issn : 0002-9262, 1476-6256. (Visité le 19/09/2016).

Alain Lefebvre 2015
Lefebvre, Alain (2015). *MAREL Carnot Data and Metadata from Coriolis Data Centre. SEANOE*. http://doi.org/10.17882/39754. (Visité le 22/11/2016).

A. Lefebvre et al. 2016
Lefebvre, A. et al. (2016). *Optimization of the monitoring strategy for the French National Phytoplankton and Phycotoxins Network (REPHY) using semi-automated digital images analysis*. International workshop on current advances in the application of (semi-)automated techniques for studying phytoplankton dynamics in coastal and marine waters. 31 mai-2 juin 2016, Wimereux, France.

Liao et al. 2014
Liao, Serena G. et al. (2014). « Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or Not, and How? » In : *BMC Bioinformatics* 15, p. 346. issn : 1471-2105. (Visité le 20/09/2016).

Lucińska et Wierzchoń 2012
Lucińska, Małgorzata et Sławomir T Wierzchoń (2012). « Spectral clustering based on k-nearest neighbor graph ». In : *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, p. 254–265.

Luxburg 2007
Luxburg, Ulrike von (2007). « A Tutorial on Spectral Clustering ». In : *CoRR* abs/0711.0189. arXiv : 0711.0189. url : http://arxiv.org/abs/0711.0189.

Moritz et Bartz-Beielstein 2017
Moritz, Steffen et Thomas Bartz-Beielstein (2017). « imputeTS: Time Series Missing Value Imputation in R ». In : *The R Journal* 9.1, p. 207–218. doi : 10.32614/RJ-2017-009. url : https://doi.org/10.32614/RJ-2017-009.

Moritz, Sardá et al. 2015
Moritz, Steffen, Alexis Sardá et al. (2015). « Comparison of Different Methods for Univariate Time Series Imputation in R ». In : *arXiv preprint arXiv:1510.03924*. (Visité le 19/09/2016).

Ng, Jordan et Weiss 2002
Ng, Andrew Y, Michael I Jordan et Yair Weiss (2002). « On spectral clustering: Analysis and an algorithm ». In : *Advances in neural information processing systems*, p. 849–856.

P.A. Hébert, **E. Caillault Poisson** et Hamad 2011
P.A. Hébert, G. Wacquet an, **E. Caillault Poisson** et D. Hamad (2011). « Semi-supervised K-way Spectral Clustering using Pairwise Constraints ». In : *NCTA 2011 - Proceedings of the International Conference on Neural Computation Theory and Applications [part of the International Joint Conference on Computational Intelligence IJCCI 2011], Paris, France, 24-26 October, 2011*, p. 72–81.

Phan 2018
Phan, Thi-Thu-Hong (2018). « Elastic matching for classification and modelisation of incomplete time series ». Thèse de doct. Université du Littoral Côte d'Opale.

Poisson Caillault et Hébert 2013
Poisson Caillault, **Emilie** et Pierre-Alexandre Hébert (2013). *R-tools : phytoplankton classification by flow cytometry and spectral fluorometry.* Atelier Pelagos RESOMAR 2013 du 4 au 6 décembre à Wimereux, France. Organisatrice de la session pratique : identification du phytoplancton à partir de l'analyse de données fluorométriques et cytométriques par des techniques de classification semi-automatisées. Formation aux outils développés en R et analyse d'échantillons marins.

Raghunathan, Lepkowski et al. 2001
Raghunathan, Trivellore E., James M. Lepkowski et al. (2001). « A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models ». In : *Survey methodology* 27.1, p. 85–96. (Visité le 19/09/2016).

Raghunathan et Siscovick 1996
Raghunathan, Trivellore E. et David S. Siscovick (1996). « A Multiple-Imputation Analysis of a Case-Control Study of the Risk of Primary Cardiac Arrest Among Pharmacologically Treated Hypertensives on JSTOR ». In : *Royal Statistical Society. Series C (Applied Statistics)* 45, p. 335–352.

Rahman et al. 2015
Rahman, Shah Atiqur et al. (2015). « Combining Fourier and Lagged K-Nearest Neighbor Imputation for Biomedical Time Series Data ». en. In : *Journal of Biomedical Informatics* 58, p. 198–207. issn : 15320464. (Visité le 19/09/2016).

Rand 1971 Rand, William M. (1971). « Objective Criteria for the Evaluation of Clustering Methods ». In : *Journal of the American Statistical Association* 66.336, p. 846–850. doi : 10.1080/01621459.1971.10482356. url : https://doi.org/10.1080/01621459.1971.10482356.

Royston 2007
Royston, Patrick (2007). « Multiple Imputation of Missing Values: Further Update of Ice, with an Emphasis on Interval Censoring ». In : *Stata Journal* 7.4, p. 445–464. (Visité le 19/09/2016).

Rubin 1996 Rubin, Donald B. (1996). « Multiple Imputation after 18+ Years ». In : *Journal of the American Statistical Association* 91.434, p. 473–489. doi : 10.1080/01621459.1996.10476908. url : https://doi.org/10.1080/01621459.1996.10476908.

Sakoe et Chiba 1978
Sakoe, H. et S. Chiba (1978). « Dynamic programming algorithm optimization for spoken word recognition ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1, p. 43–49. issn : 0096-3518. doi : 10.1109/TASSP.1978.1163055.

Schafer 1997
Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London : Chapman et Hall.

Scott et Varian 2014
Scott, Steven L. et Hal R. Varian (2014). « Predicting the present with Bayesian structural time series ». In : *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1/2, p. 4. doi : 10.1504/ijmmno.2014.059942. url : https://doi.org/10.1504/ijmmno.2014.059942.

Shah et al. 2014
Shah, Anoop D. et al. (2014). « Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study ». eng. In : *American Journal of Epidemiology* 179.6, p. 764–774. issn : 1476-6256.

Shumway et Stoffer 2017
Shumway, Robert H. et David S. Stoffer (2017). *Data sets and scripts to accompany Time Series Analysis and Its Applications: With R Examples (4th ed)*. Springer International Publishing. doi : 10.1007/978-3-319-52452-8. url : https://doi.org/10.1007/978-3-319-52452-8.

Spratt et al. 2010
Spratt, M. et al. (2010). « Strategies for Multiple Imputation in Longitudinal Studies ». en. In : *American Journal of Epidemiology* 172.4, p. 478–487. issn : 0002-9262, 1476-6256. (Visité le 20/09/2016).

D. J. Stekhoven et Buhlmann 2011
Stekhoven, D. J. et P. Buhlmann (2011). « MissForest–non-parametric missing value imputation for mixed-type data ». In : *Bioinformatics* 28.1, p. 112–118. doi : 10.1093/bioinformatics/btr597. url : https://doi.org/10.1093/bioinformatics/btr597.

Daniel J. Stekhoven et Bühlmann 2012
Stekhoven, Daniel J. et Peter Bühlmann (2012). « MissForest—non-parametric missing value imputation for mixed-type data ». In : *Bioinformatics* 28.1, p. 112–118. doi : 10.1093/bioinformatics/btr597.

Stuart et al. 2009
Stuart, Elizabeth A. et al. (2009). « Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative ». en. In : *American Journal of Epidemiology* 169.9, p. 1133–1139. issn : 0002-9262, 1476-6256. (Visité le 19/09/2016).

Suryanarayana, Rao et Swamy 2015
Suryanarayana, SV, G Venkateswara Rao et G Veereswara Swamy (2015). « A Survey: Spectral Clustering Applications and its Enhancements ». In : 6.

Thoning, Tans et Komhyr 1989
Thoning, Kirk W, Pieter P Tans et Walter D Komhyr (1989). « Atmospheric Carbon Dioxide at Mauna Loa Observatory. II - Analysis of the NOAA GMCC Data, 1974-1985 ». EN-US. In : 94, p. 8549–8565. (Visité le 23/09/2016).

T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018
T.T.H. Phan, A. Bigand et **E. Poisson Caillault** (2018). « A New Fuzzy Logic-Based Similarity Measure Applied to Large Gap Imputation for Uncorrelated Multivariate Time Series ». In : *Applied Computational Intelligence and Soft Computing* 2018, 9095683:1–9095683:15. doi : 10.1155/2018/9095683.

T.T.H. Phan, **E. Poisson Caillault** et Bigand 2016
T.T.H. Phan, **E. Poisson Caillault** et A. Bigand (2016). « Comparative study on supervised learning methods for identifying phytoplankton species ». In : *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)*, p. 283–288. doi : 10.1109/CCE.2016.7562650.

T.T.H Phan, **E. Poisson Caillault** et Bigand 2019
T.T.H Phan, **E. Poisson Caillault** et A. Bigand (2019). « eDTWBI: effective imputation method for univariate time series ». In : *International Conference on Computer Science, Applied Mathematics and Applications - ICCSAMA'2019, Hanoi, Vietnam.*

T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre et al. 2017
T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre et al. (2017). « Dynamic time warping-based imputation for univariate time series data ». In : *Pattern Recognition Letters (IF: 1.952)*. doi : 10.1016/j.patrec.2017.08.019.

T.T.H. Phan, **E. Poisson-Caillault** et Bigand 2018
T.T.H. Phan, **E. Poisson-Caillault** et A. Bigand (2018). « Comparative Study on Univariate Forecasting Methods for Meteorological Time Series ». In : *2018 26th European Signal Processing Conference (EUSIPCO)*, p. 2380–2384. doi : 10.23919/EUSIPCO.2018.8553576.

T.T.H Phan et al. 2017
T.T.H Phan et al. (2017). « DTW-Approach for uncorrelated multivariate time series imputation ». In : *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, p. 1–6. doi : 10.1109/MLSP.2017.8168165.

Van Buuren, Boshuizen, Knook et al. 1999
Van Buuren, Stef, Hendriek C. Boshuizen, Dick L. Knook et al. (1999). « Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis ». In : *Statistics in medicine* 18.6, p. 681–694. (Visité le 19/09/2016).

Wacquet et al. 2018
Wacquet, G. et al. (2018). *Combination of machine learning methodologies and automated data acquisition systems for phytoplankton detection and classification*. AG JERICO NEXT, Galway.

Wagner, Miller et Garibaldi 2011
Wagner, Christian, Simon Miller et Jonathan M. Garibaldi (2011). « A fuzzy toolbox for the R programming language ». In : *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*. IEEE. doi : 10.1109/fuzzy.2011.6007743. url : https://doi.org/10.1109/fuzzy.2011.6007743.

Wagstaff, Cardie et al. 2001
Wagstaff, Kiri, Claire Cardie et al. (2001). « Constrained K-means Clustering with Background Knowledge ». In : *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 577–584.

Wagstaff, Davidson et Basu 2008
Wagstaff, Kiri, Ian Davidson et Sugato Basu, éds. (2008). *Constrained Clustering*. Chapman et Hall/CRC. doi : 10.1201/9781584889977. url : https://doi.org/10.1201/9781584889977.

Walter.O et al. 2013
Walter.O, Yodah et al. (2013). « Imputation of Incomplete Non- Stationary Seasonal Time Series Data ». en. In : *Mathematical Theory and Modeling* 3.12, p. 142–154. issn : 2225-0522. (Visité le 22/09/2016).

Wang et Davidson 2010
Wang, Xiang et Ian Davidson (2010). « Flexible constrained spectral clustering ». In : *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 563–572.

Xie et Wiltgen 2010
Xie, Ying et Bryan Wiltgen (2010). « Adaptive Feature Based Dynamic Time Warping ». In : *IJCSNS International Journal of Computer Science and Network Security* 10.1.

Yan, Huang et Jordan 2009
Yan, Donghui, Ling Huang et Michael I Jordan (2009). « Fast approximate spectral clustering ». In : *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 907–916.

Zeileis et Grothendieck 2005
Zeileis, Achim et Gabor Grothendieck (2005). *zoo: S3 Infrastructure for Regular and Irregular Time Series, used package in 2016*. doi : 10.18637/jss.v014.i06. url : https://www.jstatsoft.org/v014/i06.

Zelnik-Manor et Perona 2005
Zelnik-Manor, Lihi et Pietro Perona (2005). « Self-tuning spectral clustering ». In : *Advances in neural information processing systems*, p. 1601–1608.

Chapitre 3

Segmentation de séries temporelles

« Écrire ou segmenter sa pensée,
nécessite d'identifier
ce qui peut paraître percutant. »

EPC

Sommaire

3.1 Introduction **58**
3.2 Segmentation guidée par les données **62**
3.2.1 Segmentation par classification non supervisée 62
3.2.2 Modélisation par Modèle de Markov Caché hybride non supervisé . . 62
3.3 Intégration des connaissances temporelles **68**
3.3.1 Classification guidée par des contraintes temporelles 68
3.3.2 Contraintes déterminées par la phénologie 70
3.4 Segmentation d'événements ou zones isolés ou atypiques **72**
3.5 Conclusions, Perspectives. **73**

Ce chapitre résume des travaux liés à la de la segmentation associés aux publications principales indiquées :

— Une brève présentation des difficultés de segmentation des séries temporelles et des travaux existants ;

— La présentation de la méthodologie pour obtenir une segmentation non supervisée en événements intermittents basée sur une classification spectrale. (**E. Poisson Caillault** et Ternynck 2013 ; Lefebvre et **E. Poisson Caillault** 2019 ; K. Rousseeuw, **E. Caillault** et al. 2016 ; Lefebvre, **E. Poisson-Caillault** et al. 2016) ;

— La proposition d'un hybride markovien pour modéliser une série temporelle multivariée (K. Rousseeuw, **E. Poisson Caillault** et al. 2013 ; K. Rousseeuw, **E. Poisson Caillault** et al. 2015) ;

— La proposition d'un classifieur spectral divisif pour identifier des événements rares ou extrêmes (K. Grassi, **E. Poisson Caillault** et Lefebvre 2019) ;

— Une ouverture à l'insertion de connaissance temporelle : proposition d'une méthode de classification spectrale contrainte temporellement et une étude phénologique d'une série segmentée par recherche de mélange de courbes pour améliorer les paramètres dynamiques du modèle de Markov Caché (**E. Poisson Caillault** et Lefebvre 2017).

3.1 Introduction

"La vie est une succession d'instants et de rencontres que seule la photographie a le pouvoir d'immortaliser" (Szczepan Yamenski). Bien que les technologies actuelles permettent d'obtenir cette "photographie" d'une variable ou d'un processus à des échelles spatiales et temporelles riches, identifier ses instants caractéristiques et comprendre leurs formations et conséquences nécessitent des outils adaptés selon la nature de la variable et de ses instants. Reprenons l'exemple de la fluorescence mesurée en zone côtière comme proxy de la biomasse phytoplanctonique, c'est un processus non stationnaire dont les années présentent une variabilité forte.

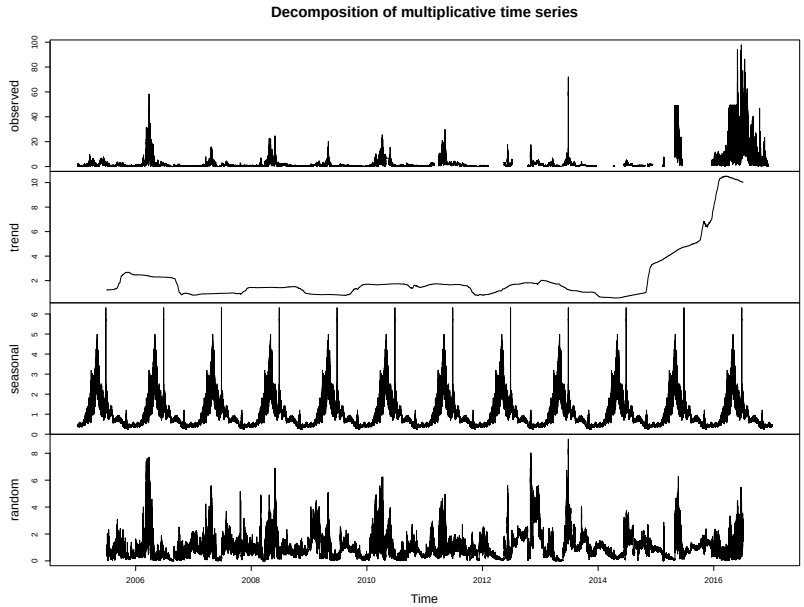


FIGURE 3.1 – Visualisation de la décomposition multiplicative tendance-cycle du signal de fluorescence mesurée par la station MAREL-Carnot (données Coriolis, IFREMER).

La figure 3.1, retraçant les observations de fluorescence et sa décomposition tendance-cycle au fil des années pour la station MAREL-Carnot à Boulogne-sur-mer, et la figure 3.2 correspondant aux distributions journalières toute année confondue, illustre bien cette notion de variabilité en amplitude et temporelle.

En effet, les résidus loin d'être du bruit, présentent des informations indispensables pour détecter les efflorescences phytoplanctoniques dont certaines ne sont pas visibles dans le cycle de saisonnalité obtenu par décomposition. Et, ce cycle annuel (tracé rouge ou bleu pour une décomposition respectivement additive ou multiplicative) superposé à la figure 3.2 ne reflète pas l'ensemble des pics liés aux successions des efflorescences printanières, estivales et automnales, qui possèdent des irrégularités tant en amplitude que dans leur datation et durée. Identifier les périodes phytoplanctoniques, productive et non productive, peut paraître assez évident : la stratégie directive cadre marine (DCSMM) et celle cadre sur l'eau (DCE) préconise d'intensifier les prélèvements entre mars et octobre, date dite de la période productive. Par des approches de classification conventionnelle, il est facile à minima de différencier ces deux périodes. La figure 3.3 illustre le résultat d'une classification kmeans à partir de 9 paramètres mesurés EOV entre 2005 et 2010 avec K fixé à 2 pour les 2 périodes souhaitée. Certes, un fort chevauchement et un étalement de ces deux états sont présents. Le recouvrement est expliqué par une phénologie des trois efflorescences très variées et étalées toutes années confondues (cf. figure 3.2).

Identifier ensuite les périodes de production ou apparition d'algues nuisibles ou toxiques devient un challenge plus difficile nécessitant des développements méthodologiques et numé-

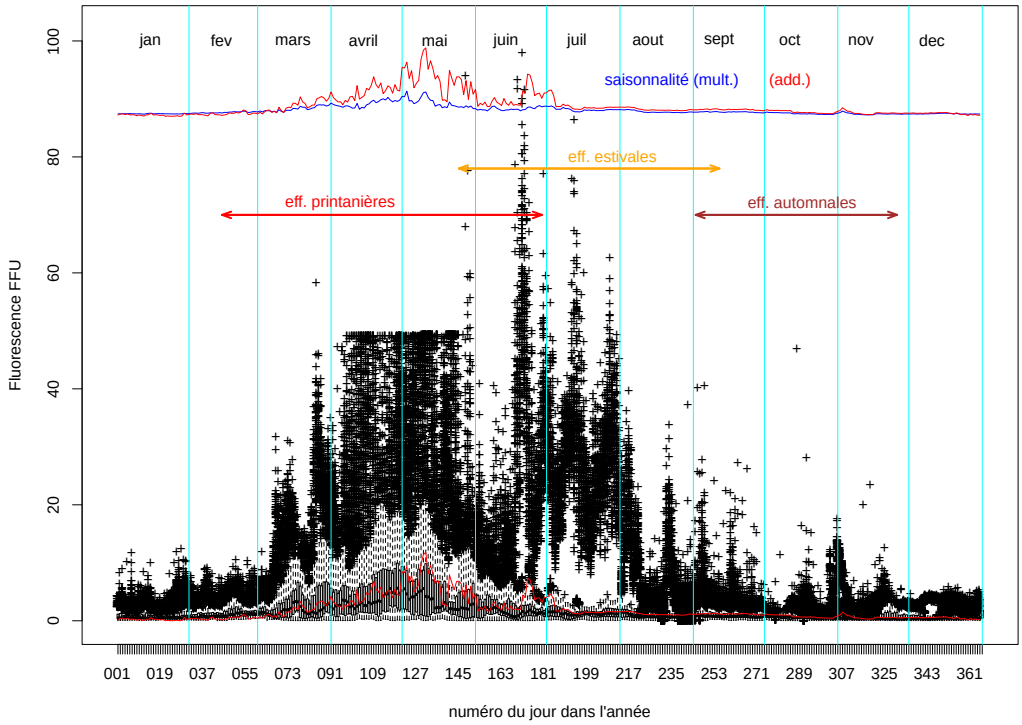


FIGURE 3.2 – Visualisation des variabilités journalières des observations de fluorescence mesurée par la station MAREL-Carnot (données Coriolis, IFREMER).

riques non conventionnels. En effet, les prélèvements et analyses associées en flux nécessitent des experts taxonomiques (telle la microscopie, la cytométrie pour différencier des groupes phytoplanctoniques ou la fluorescence multispectrale pour différencier des classes pigmentaires) et un temps d'études ne permettant ni de discriminer potentiellement toutes les classes nuisibles ni de traiter l'ensemble des prélèvements effectués en infra-journalier.

À l'heure actuelle, nous ne disposons pas des connaissances nécessaires pour interpréter les distributions à haute résolution temporelle et/ou spatiale des efflorescences par espèce phytoplanctonique. Les travaux associés dans la littérature sur la recherche de motifs connus est donc inexploitable en l'état sans supposition. Une voie est la recherche de points de rupture caractéristiques. Ces points de coupures sont obtenus à partir d'un critère de dissemblance entre deux segments adjacents tels les plus élémentaires : la moyenne, la variance, moyenne et variance ou le spectre.

Nous présentons le résultat de trois algorithmes de détection de ruptures dans un jeu artificiel sur la figure 3.4 puis sur le jeu MAREL-Carnot avec un zoom sur l'année 2005. La série-jeu artificielle composée de trois signaux $Y = \{y_1, y_2, y_3\}$ de taille T=1 000 observations a été construite avec deux événements extrêmes notés ev1 et ev2 et un événement de type anomalie capteur ev3 à l'intérieur d'un signal saisonnier (noté gs : global shape).

La première ligne correspond à la segmentation obtenue à partir d'une technique de détection des ruptures en moyenne et en variance dans une série univariée (signal y_3 pour le jeu artificiel et le signal de fluorescence pour MAREL-Carnot). Cette approche est basée sur une fonction de coût optimisant conjointement le nombre de ruptures et leurs positions (KILLICK, FEARNHEAD et I. A. ECKLEY 2012 ; KILLICK et I. ECKLEY 2014). Neuf points de rupture sont obtenus sur le signal y_3 liés selon la position du bruit inséré, les segments obtenus ne permettent pas d'isoler les événements introduits. 2 294 points de rupture sont obtenus pour le signal de fluorescence MAREL-Carnot, seuls ceux de l'année 2015 sont représentés. Cette sur-segmentation sera difficile à interpréter par un expert même après clustering pour

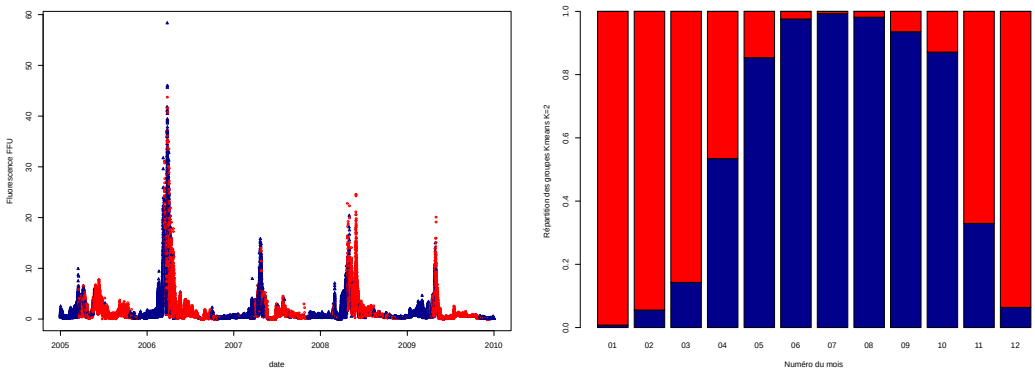


FIGURE 3.3 – Partitionnement k-means en deux groupes des données prétraitées MAREL-Carnot : Projection des clusters sur le signal de fluorescence mesurée par la station MAREL-Carnot de 2005 à 2010 et à gauche la répartition mensuelle toutes années confondues.

réduire la quantité d'information à labelliser.

La seconde approche est basée sur une approche hiérarchique divisive avec une distance Energy-statistic (SZEKELY et RIZZO 2013). Cette approche permet d'avoir une représentation de type arbre binaire. La racine de cet arbre correspond à la série entière, celle-ci sera coupée en deux feuilles (sous-séquences) selon une distance maximale entre les distributions multivariées de segments adjacents et ainsi de suite pour les niveaux inférieurs. Dans le cas artificiel, cette approche isole les trois événements attendus mais sur-segmente le signal global (gs). Seulement 13 ruptures au total, dont 2 (et celle à t=0) pour 2005, sont obtenues pour le signal de fluorescence. Celles-ci ne correspondent pas à l'isolement de pics particuliers mais à des coupures sur les maxima de certains pics.

La dernière approche illustrée est une technique de classification agglomérative. Initialement chaque observation est associée à son propre segment de longueur 1. Les segments voisins sont séquentiellement regroupés s'ils optimisent un critère de dissemblance basé sur la mesure de divergence des distributions entre segments voisins. Les résultats sur le signal y_3 isolent les deux premiers événements mais ne permet pas de distinguer le troisième d'une portion du signal global. Pour la fluorescence, 2 790 points de ruptures sont obtenus entraînant une sur-segmentation. A leur décharge, les auteurs de cet algorithme précisent que cette méthode requièrent une segmentation initiale pour être efficace.

Afin d'apporter une aide à la labellisation par un expert humain en états environnementaux de ces séries, nous avons fait le choix de travailler sur la géométrie des données et donc de segmenter les données par des outils de classification non supervisée. Ainsi en n'incluant pas l'information temporelle (ou spatio-temporelle) directement dans notre recherche, deux événements sur- ou sous-représentés à deux dates/années éloignées pourront être catégorisés ensemble si les signaux correspondants présentent les mêmes caractéristiques.

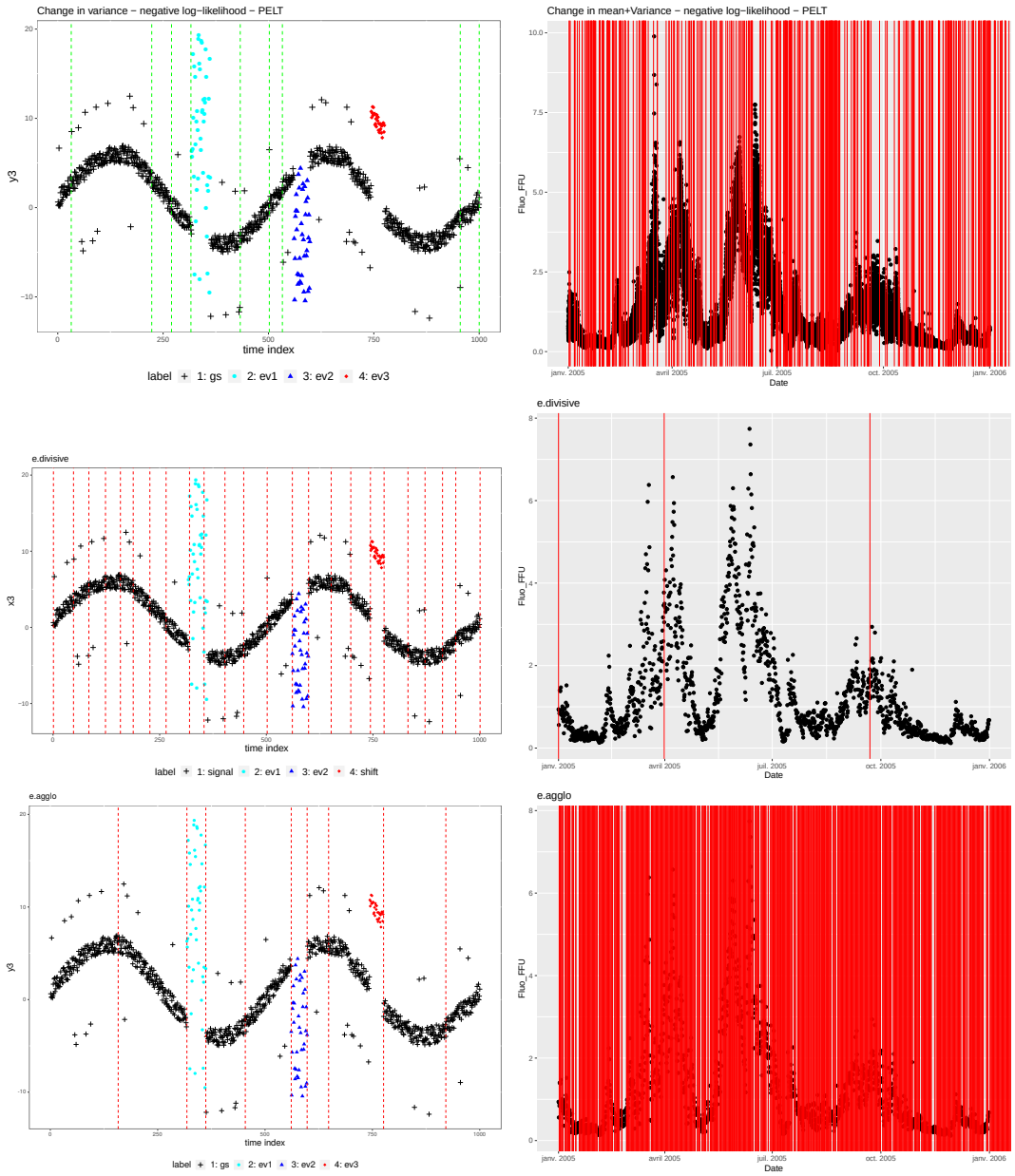


FIGURE 3.4 – Segmentation par des approches de détection de changement de variance dans un jeu de données artificiel (à droite) et les données MAREL Carnot (à gauche). De haut en bas : Segmentation obtenue obtenue à partir des changements en moyenne et variance dans le signal y_3, Points de coupure par approche divisive multivariée, puis par approche agglomérative multivariée.

3.2 Segmentation guidée par les données

Une série est une séquence d'événements S_t de N variables observées, $S_t \in \{s_1, \ldots, s_N\}$ qui peut être représentée comme un graphe orienté avec une dynamique associée notamment : la probabilité de naviguer entre événements ou de rester dans un événement ; probabilité d'émission d'un événement. Ce graphe ergodique n'est pas une séquence explicitement linéaire. La représentation choisie correspond à une formalisation par Modèle de Markov Caché (MMC) $\lambda(N = |S|, M, A, B, \pi)$ avec N le nombre d'événements (états environnementaux - S pour State), M le nombre de symboles pour caractériser ces événements, $A(N \times N)$ la matrice de transition entre événements, $B(N \times M)$ la matrice des probabilités d'émission, $\pi = \{\pi_i, i \in [1, \ldots, N]\}$ le vecteur associé aux probabilités de commencer par tel événement (Rabiner 1989).

3.2.1 Segmentation par classification non supervisée

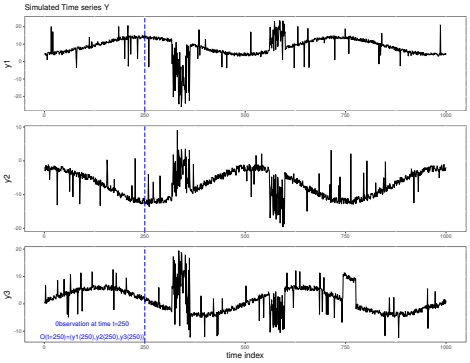
Pour obtenir cette vue basée événements, à chaque observation O_t est assigné un label événements s_i. Nous avons fait le choix d'une labellisation par une technique de classification spectrale n'imposant aucune hypothèse sur la distribution et la forme de ces événements. L'ensemble des observations fournies sont ainsi traduites sous la forme d'un graphe orienté où les poids des arcs w_{ij} correspondent à une mesure de proximité (géométrique et/ou temporelle) entre chaque observation i et j. La figure 3.5 reprend le processus de segmentation puis de modélisation du jeu artificiel $Y = y_1, y_2, y_3$. Afin de bien comprendre le processus de clustering, nous nous plaçons dans un cadre idéal où la matrice de similarité et son laplacien sont bloc-diagonales induisant une estimation parfaite par ses valeurs propres et vecteurs propres. La matrice de similarité W, normalement calculée à partir de la seule information $Y = y_1, y_2, y_3$, est ici construite à partir des labels : $w_{ij} = W(O_i, O_j) = 1$ pour deux observations de même label, 0 sinon. W étant idéale, bloc-diagonale, le nombre de valeurs propres égales à un correspond au nombre K de clusters souhaité et les observations appartenant à un même label dans l'espace spectral normalisé (des K vecteurs propres extraits du Laplacien $L_{NJW} = D^{-1/2} W D^{-1/2}$) sont idéalement regroupées en un unique point sur la sphère unité.

3.2.2 Modélisation par Modèle de Markov Caché hybride non supervisé

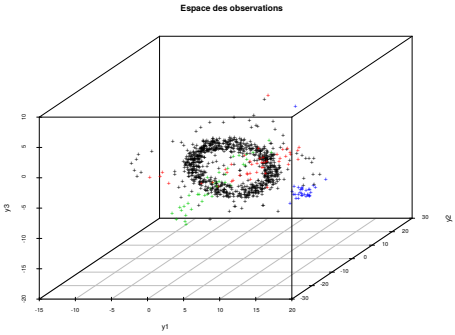
La classification spectrale fournit ainsi une estimation N du nombre d'états-événements d'un modèle HMM d'ordre 1 $\lambda(N = |S|, M, A, B, \pi)$ par la méthode des valeurs propres dominantes ou du maximum de gap entre les valeurs propres successives (G. Wacquet, **E. Poisson-Caillault** et Hebert 2013). Elle fournit un label état à chaque observation, permettant ainsi de calculer directement les probabilités de passage d'un état à un autre soit la matrice de transition $A = \{a_{ij} = P(s_j(t) \mid s_i(t-1)))\}$. Sans *a priori*, le vecteur de probabilités initiales π est fixé de manière équiprobable entre chaque état ainsi $\pi = \{P(s_i) = 1/N\}$.

Dans un cadre élémentaire, l'observation des symboles d'un HMM correspond aux sorties du système ou une loi d'émission souvent gaussienne ou des lois non elliptiques. Les événements tels les blooms phytoplanctoniques sont des processus non stationnaires de formes très variables. De plus, ils ne se caractérisent pas par un vecteur de paramètres physiques unique. Nous avons donc fait le choix de réaliser une quantification vectorielle de notre espace pour construire l'ensemble de M symboles. A partir de cette réduction de l'espace en symboles, chaque observation est labellisée comme appartenant à un symbole v et un état s particulier $(O(t), v_j, s_i)$. La matrice d'émission est alors calculée directement sur cette base double étiquetée $B_{ij} = P(v_j|s_i)$.

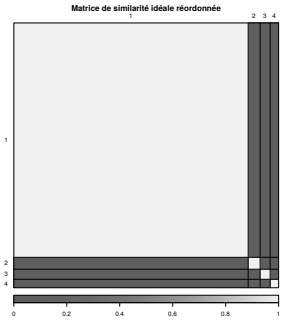
Ainsi, nous obtenons un modèle hybride SC-HMM (Spectral Clustering - Hidden Markov Model) s'affranchissant des hypothèses usuelles sur N et/ou M et des étapes de maximisation des paramètres de vraisemblance des observations et de l'état à partir d'un modèle donné, fixé initialement. Par conséquent, les paramètres probabilistes (A, B, π) sont calculés sans



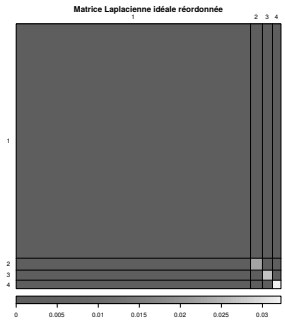
(a) Représentation temporelle



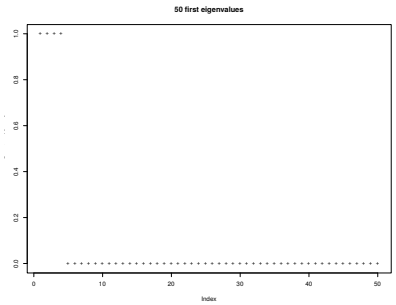
(b) Espace des variables



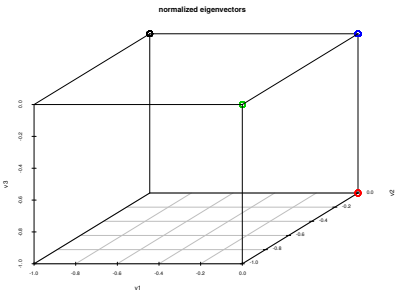
(c) Image de la matrice de similarité W



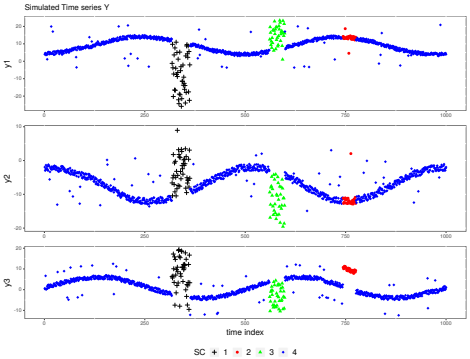
(d) Image du laplacien L_{NJW}



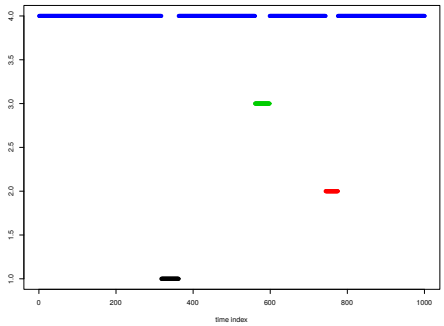
(e) Valeurs propres de L_{NJW}



(f) Espace spectral et clustering



(g) Série segmentée par SC



(h) Dynamique des labels

FIGURE 3.5 – Processus de segmentation par classification dans l'espace spectral extrait des variables des observations.

processus itératif.

Dans K. Rousseeuw, **E. Poisson Caillault** et al. 2015, une optimisation de ce processus a été réalisée en extrayant les états directement à partir des symboles obtenus sur un critère de réduction par quantification conservant un minimum de 95 % de la variance expliquée des données. Le principe est illustré à la figure 3.6.

L'algorithme 4 de réduction nommé STFKM pour Self Tuning Fast K-Means est basé sur un critère de variance expliquée avec MS - Mean Square - l'erreur moyenne au carré de la variance inter, ou intra ou totale K. Rousseeuw, **E. Poisson Caillault** et al. 2013).

Algorithme 4 Algorithme (non traduit) STFKM - Self Tuning Fast K-Means algorithm. MS = Mean Square of variance (between, within or total)

Require: $\mathbf{O}, Kmax, varExplained$
1: **if** $varExplained$ not defined **then**
2: $varExplained$=0.95
3: **end if**
4: **if** $Kmax$ not defined **then**
5: $Kmax$=nrow($\mathbf{O}$)
6: **end if**
7: Variable : k=1, vE=0 ;
8: **while** $k < Kmax$ or $vE < varExplained$ **do**
9: $k = k + 1$;
10: Step 1 : Initialization of k centers
11: ..Cut Data in n subsamples of 20,000 points
12: ..Compute K-means (K=k) on each subsample
13: ..Select the k clusters centers from the best partition according $MS(within)/MS(between)$
14: Step 2 : Decide the class memberships of the N_p points by assigning them to the nearest center.
15: Step 3 : Re-estimate the k cluster centers, by assuming the new memberships found
16: Step 4 : If none of N_p points changed membership in the last iteration, Otherwise goto 2.
17: Step 5 : $vE = MS(between)/MS(total)$
18: **end while**
19: **return** k obtained centers

La thèse de Kevin Rousseeuw co-encadrée avec Alain Lefebvre (expert-sénior IFREMER LER-BL) reprend différents scénarios de validation de cet hybride et de comparaison avec des approches supervisées et non supervisées (Kévin Rousseeuw 2014).

Expérimentations numériques

L'algorithme de construction de cet hybride appelé uHMM (pour "unsupervised HMM") a fait l'objet d'un transfert vers la communauté scientifique. Un package R-CRAN (https://cran.r-project.org/web/packages/uHMM/index.html) a été déposé intégrant l'ensemble des fonctions mais aussi une interface permettant à des experts et non-experts de tester leur série de données. Une amélioration du processus de construction des symboles a été ajouté afin de mieux respecter la notion de voisinage induite par l'utilisation de la similarité de Zelnick et Perona. Par quantification vectorielle, M médoides sont retenus auquel sont ajoutés un nombre M'(M) d'observations appartenant au même cluster et tirées aléatoirement. Ce nombre est fonction de la densité du cluster M d'appartenance et de la taille maximale de symboles autorisées (certes, fonction de la capacité du processeur utilisé - actuellement $max(M') = 11$ et $max(M + M') = 2500$). La longueur des séries étant courtes, les étapes de construction des symboles et des états sont indépendantes.

Plusieurs formations et ateliers pratiques ont été menés dans le cadre des journées RESOMAR, projet INTERREG IVa 2 mers DYMAPHY, projet H2020 JERICO-NEXT pour le contexte d'évaluation de la qualité des eaux côtières et larges mais aussi pour des applications à visée terrestre comme le suivi physico-chimique et algal du Marais de Saint-Quentin (collaboration CEREMA : Prygiel et al. 2018), ou la mesure de l'impact de la navigation sur la Deule (collaboration/contrat Agence de l'Eau Artois Picardie).

La figure suivante 3.7 illustre la dynamique des états obtenus sur les données MAREL-Carnot. La segmentation est réalisée à partir de la série constituée de 9 paramètres EOV sur les années 2005-2008. La fluorescence, considérée comme la variable à expliquer, n'a pas participé à l'étape de segmentation. Elle est utilisée uniquement pour valider l'interprétation des clusters obtenus et démontrer l'aide apportée aux experts en eutrophisation qui requièrent à la fois des outils de compréhension et de prédiction des états environnementaux et des outils d'alerte d'apparition de nouveaux phénomènes ou de défaillances capteur (cas de l'état S6 non détecté après une vérification des gammes capteurs).

Par analogie, cet outil uHMM a été utilisé pour réaliser un découpage des communautés phytoplanctoniques en Manche Mer du Nord lors d'une campagne DYPHYMA (Lefebvre et **E. Poisson Caillault** 2019). Le système, sans prise en compte des positions géographiques ni des dates d'acquisitions, a été capable d'apporter une cartographie cohérente des différents types de masse d'eau sur des acquisitions de courtes durées (LEG1 un jour, LEG2 trois jours et LEG3 une semaine). La classification a été réalisée individuellement sur chaque LEG puis sur la fusion des trois à partir de différentes combinaisons de paramètres parmi 4 concentrations algales mesurées (Green ; Blue-Green ; Brown ; Cryptophyceae) et 4 paramètres physico-chimiques (température, salinité, concentration et saturation en oxygène).

Les résultats ont été soumis à une expertise humaine. Ne sont présentées ici que les conclusions générales illustrées à la figure 3.8 relatives à la partition des masses d'eau obtenues sur la fusion des LEGS. Le nombre de cluster obtenu est de $K = 9$ par la méthode des valeurs propres dominantes et $K = 19$ par la méthode du gap $K = 18 =$ (le label cluster 0 correspond aux données non classées dues à des valeurs absentes). La différence du nombre de groupes identifiés entre les eaux anglaises et celles françaises est nette. Les eaux anglaises sont plus hétérogènes que celles des côtes françaises. Les eaux dites du large sont relativement bien isolées.

Comparée à un étiquetage réalisé à partir de l'ensemble des 6 possibilités de compositions algales réellement présentes, une approche SC à K=6 fixé et K>=6 a été capable de les identifier pleinement contrairement à des algorithmes de base (KM et classification hiérarchique Ward2). De plus, cette technique a permis d'une part de bien mettre en évidence les épisodes de blooms de l'algue nuisible *Phaeocystis Globosa* sans connaissance de son empreinte spécifique, l'analyseur algal AOA étant paramétré par défaut (l'AOA - Algal Online Analyzer - utilisé est capable de détecter au plus 4 empreintes). D'autre part, il permet d'apporter une cartographie plus fine ou différente de l'approche usuelle d'un expert fondée sur des technologies et des résolutions passées et devient un outil intéressant pour affiner la stratégie d'échantillonnage des campagnes de mesure en cours ou futures.

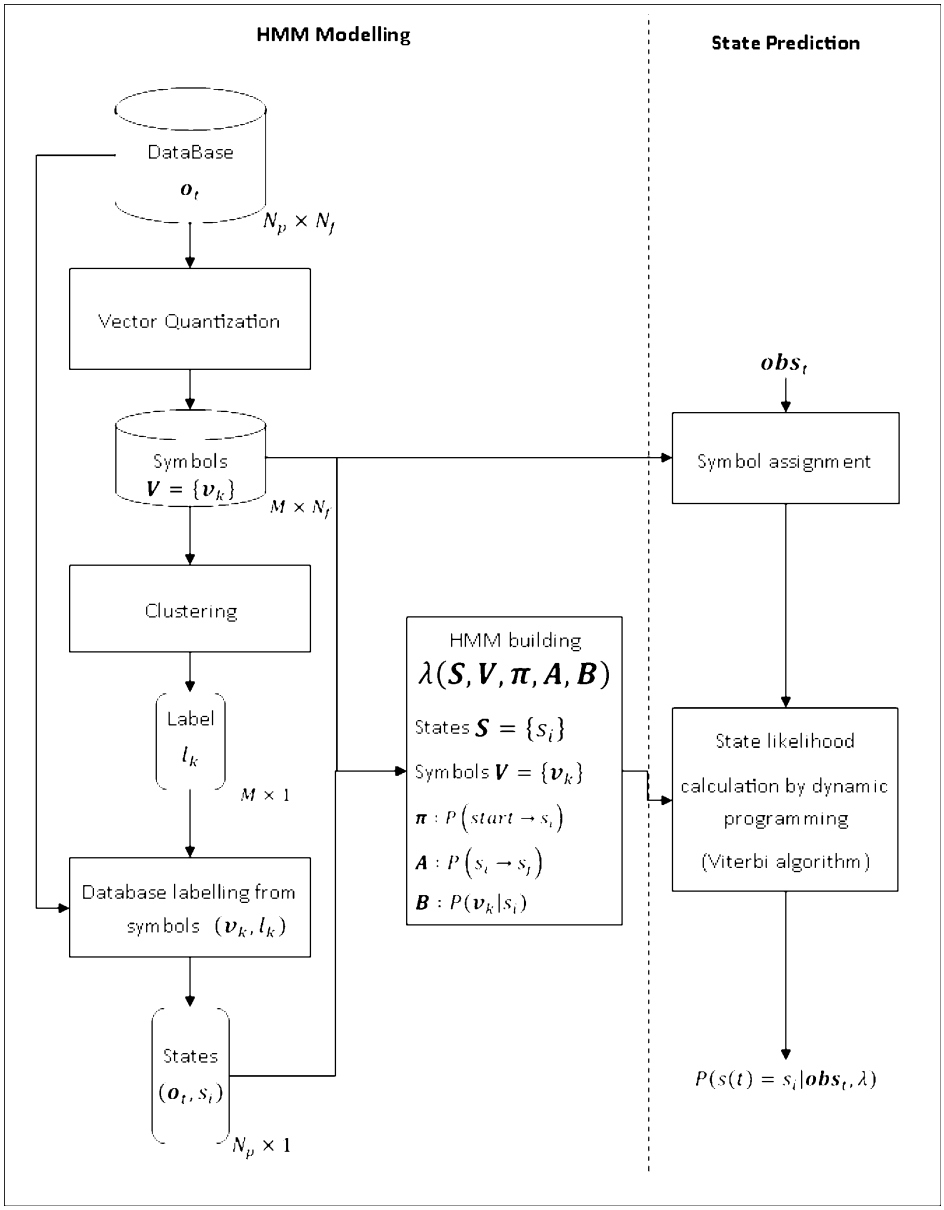


FIGURE 3.6 – Construction d'un hybride SC-HMM avec une extraction des états basées sur l'extraction des symboles.

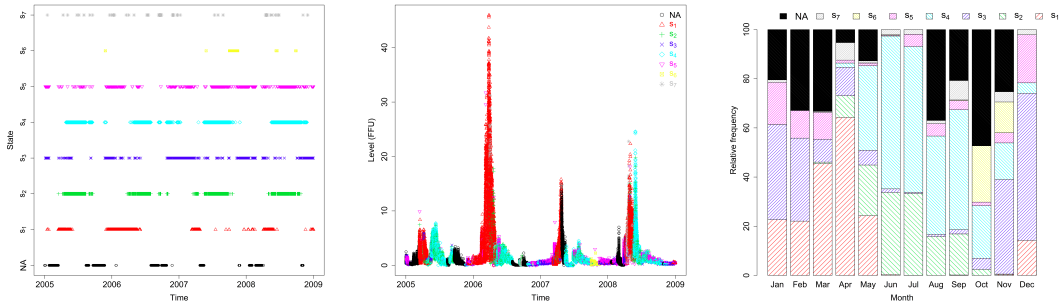


FIGURE 3.7 – Visualisation du clustering obtenu par SC-uHMM : dynamique des labels de 2005 à 2008, projection colorée des labels sur la fluorescence, répartition mensuelle des labels

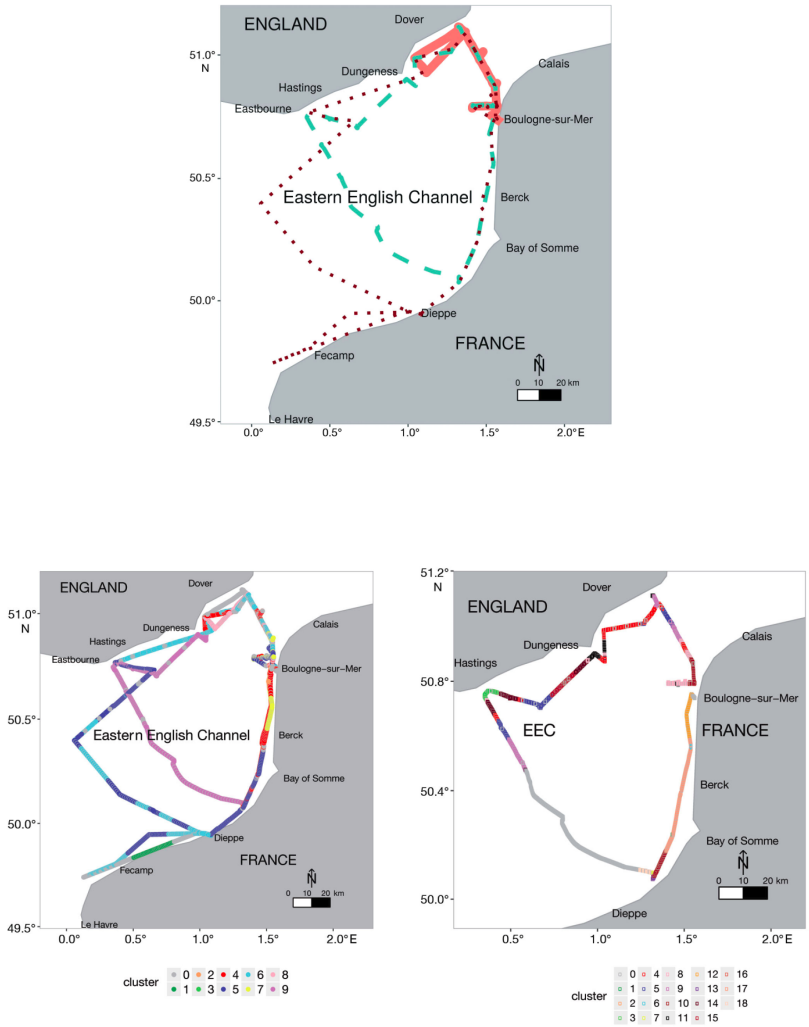


FIGURE 3.8 – Partitionnement des masses d'eau lors de legs de la campagne DYPHYMA : en haut la visualisation des trajets (LEG 1 à 3) ensuite les partitionnements par SC sur ces legs, chaque couleur désignant un label obtenu.

3.3 Intégration des connaissances temporelles

L'intégration directe de la date d'acquisition ou des coordonnées spatiales structure fortement la segmentation obtenue. Les efflorescences printanières d'une année à l'autre ou les épisodes de bloom d'une espèce d'une côte à l'autre seront forcément attribués à un état différent et une étape de comparaison des sous-séquences obtenues devra alors être mise en oeuvre pour les réassigner au même état. Pour éviter cette étape supplémentaire, une approche globale est préférée.

3.3.1 Classification guidée par des contraintes temporelles

Nous avons introduit une nouvelle méthode pour intégrer les informations temporelles et/ou spatiales en ré-utilisant l'algorithme cSC -constrained Spectral Clustering G. Wacquet, **E. Caillault Poisson** et al. 2013. La proximité géographique ou temporelle peut être transcrite sous forme de contrainte MustLink/CanNotLink. Nous formalisons ceci dans la matrice Q de l'algorithme donné au chapitre 2.

La première formulation naturelle est de considérer la connaissance sur l'évolution d'un processus. Pour reprendre les changements de communautés phytoplanctoniques, la littérature nous donne un rayon de 1 kilomètre et une durée infra-hebdomadaire. La matrice Q devant tenir compte de cette connaissance tout en conservant une certaine incertitude sur la transition liée à la variabilité de commencement, de fin et de maintien des phénomènes, nous avons choisi d'utiliser une fonction alpha de forme trapézoïdale ou triangulaire pour construire la matrice de contrainte temporelle $Q_{ij} = Q(O_i, O_j)$ est défini par :

$$Q_{ij} = \begin{cases} 0 & \text{si } |i-j| \geq T \\ \text{alpha}(i,j,T) & \text{sinon} \end{cases}$$

dans un cadre Must-Link. De façon plus discriminante, le zéro sera substitué par une contrainte CanNotLink avec une valeur à -1.

Le choix de la fenêtre T dépendra de l'application et de la connaissance sur la vitesse d'évolution du procédé. La construction de la valeur d'affinité temporelle (et/ou spatiale) est illustrée dans le cas d'une série non trouée puis d'une série avec des données manquantes (NA).

Expérimentations numériques

La figure 3.10 met en avant les différences de segmentations obtenues entre une approche spectrale classique et une approche spectrale par contrainte temporelle triangulaire avec un nombre de cluster fixé (K=6).

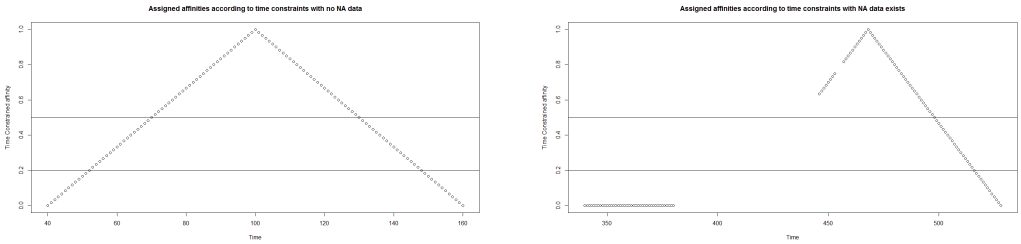


Figure 3.9 – Exemple de valeur Q_i pour une fenêtre $T = 120$ sans ou avec données absentes.

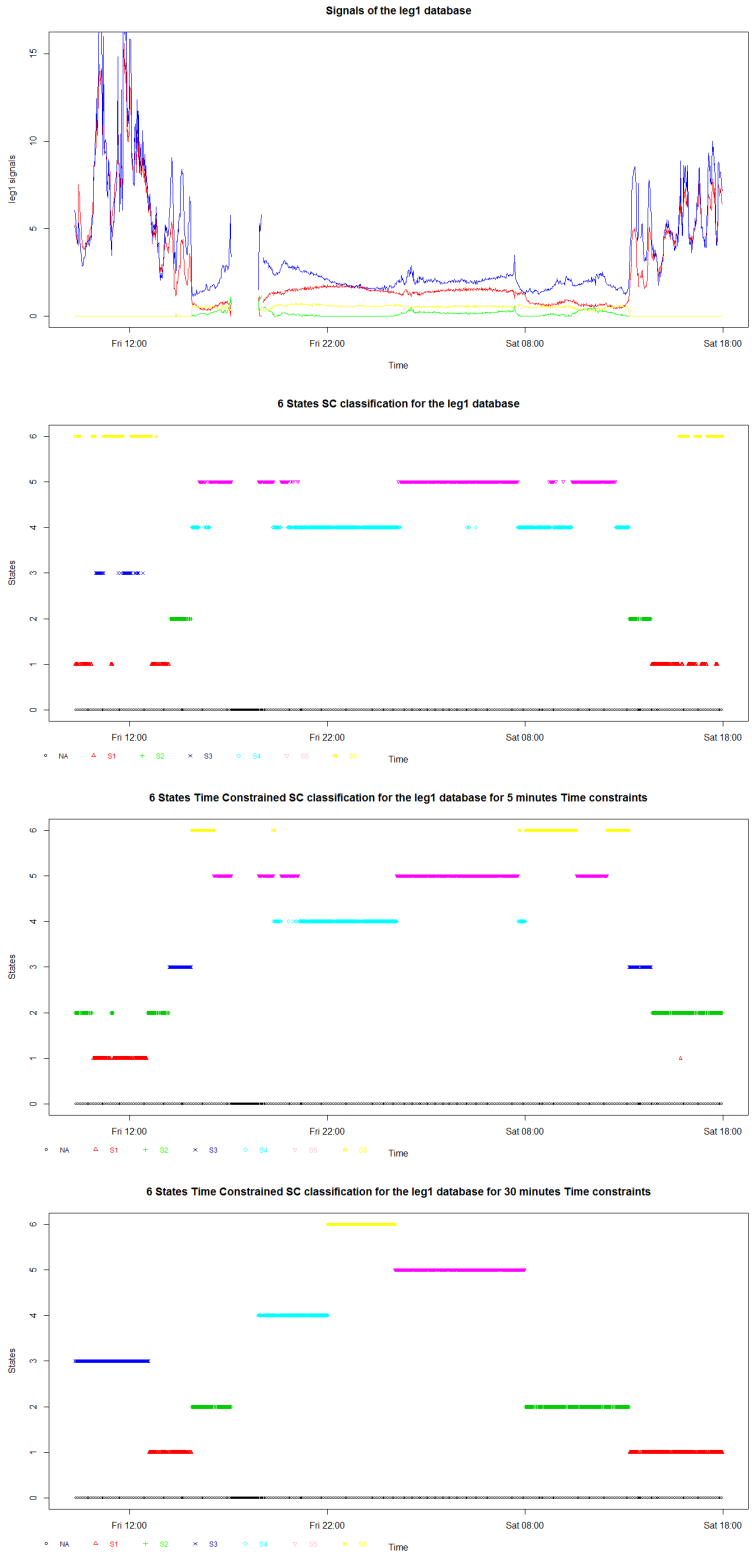


FIGURE 3.10 – Segmentation de la série formée des 4 paramètres de l'AOA signaux (en haut) du LEG 1 de la campagne DYPHYMA par clustering non contraint (image 2) et contrainte (3 et 4 en dessous)

Cette expérience a été réalisée à partir du jeu AOA (Lefebvre et **E. Poisson Caillault** 2019) sur le LEG1 de la campagne DYPHYMA ayant une fréquence d'échantillonnage de 1 minutes à mettre en corrélation avec la vitesse du bateau non constante. Les conclusions sont qu'avec un poids équivalent entre la matrice de similarité dit aussi d'affinité géométrique W et la matrice d'affinité temporelle (*i.e.* ici aussi spatiale), les épisodes de chaque cluster sont moins entrecoupés quand la fenêtre temporelle de contrainte T est large. Ce poids ML lié à l'équation $W' = (1-\beta)W + \beta Q$ et la fenêtre T reste à paramétrer en fonction de l'expertise humaine et du questionnement posé (vision globale ou vision attentive des événements rare et/ou extrêmes).

3.3.2 Contraintes déterminées par la phénologie

Une matrice d'affinité de saisonnalité peut être pareillement construite en se basant sur les numéros des mois enregistrés selon une fonction alpha trapézoïdale. Pour un mois donné, les observations seront considérées appartenant au même cluster, leurs valeurs de contraintes seront alors égale à 1 ($Q_{ij} = 1$), et affaiblies sur les mois voisins.

Une autre possibilité est d'améliorer cette matrice en recherchant le cycle saisonnier et les différents événements le composant par des modèles génératifs afin d'obtenir une datation à la fois des déclenchements, maintiens et fins. Nous avons construit un algorithme de décomposition d'un cycle saisonnier d'un processus de blooms et d'une série en général en une somme pondérée de courbes gaussiennes, $Y_t = \sum_{i=1}^{G} g_t; g_t = lambda * \exp(-(t-mu)^2/(2 * sigma^2))$. L'objectif de cet algorithme présenté dans l'encart de l'algorithme 5 est d'apprendre les paramètres de chaque courbe (mu, sigma, lambda) et du nombre de courbes par un processus itératif basé sur un critère combinant à la fois le respect de plusieurs indicateurs de bonne reconstruction et d'un recouvrement limité entre deux gaussiennes et pénalisant le nombre G de gaussiennes. Les approches basées EM n'étant pas exactes et conduisant à des solutions d'optimum locaux, le processus est réitéré T fois à chaque incrément de G. Sur les T solutions obtenues à chaque valeur de G, un seul modèle est retenu : le plus robuste et stable parmi ces T solutions.

Les indicateurs de reconstruction calculés entre Y_t et $\sum_t g_t$ choisis sont les suivants : R2.cor ; NMSE ; simArea ; FA2 ; FB ; FS ; MG ; VG avec les bornes d'acceptabilité données au chapitre 1.

Une expérimentation sur le cycle saisonnier du signal de fluorescence mesurée dans le cadre de la surveillance écologique de l'environnement côtier devant la centrale de Gravelines a permis de mettre en avant des événements reflétant bien la détection d'une catégorie de phytoplancton (telles l'algue nuisible *Phaeocystis Globosa* ou celle neurotoxique *Pseudo-nitzschia*). Ces événements marqués par des indices de diversité (Shannon) faibles sont très (dé)structurant pour l'écosystème, et peuvent même avoir des conséquences négatives pour l'homme. Toujours dans le cadre de cette étude, l'algorithme appliqué non pas sur le cycle saisonnier mais sur chaque séquence annuelle retourne un nombre de gaussiennes très variables d'une année à l'autre (du simple au double de celui du cycle) et de formes elles-mêmes très différentes. En l'état, il devient difficile d'appliquer la phénologie obtenue (datations) dans le protocole de contrainte temporelle. Ces travaux doivent être poursuivis avec pour premières pistes : la généralisation de la forme gaussienne à des formes diverses, la stabilisation du nombre de gaussiennes (variant de deux unités pour une même entrée).

Algorithme 5 PseudoCode of curve mixture detection

Require: (t, x, acceptRate)
Ensure: 3 vectors (mu, sigma, lambda)
Initialisation : test ← false ;
Initialisation : mu, sigma, lambda ← null ;
$y \leftarrow x/\max(x)$;
maxG ← compute the max accepted number of Gauss curves according to the number of peaks and valleys ;
yh ← transform y to obtain a density repartition
while test==false and $g \leq$maxG **do**
  # research of a robust model
  for $i = 1$ to 20 **do**
    mixmodel(i)←normalmixEM(yh,k= g) ;
  end for
  # Research of the best representative model
  cluster←kmeans(mixmodel, k=3)
  index←select the dominant group according the criterion of 50% mu in the same group are close. (no singleton cluster).
  for i in index **do**
    C_i ←sumGaussCurve(mixmodel(i),t)
    $score_i$ ←compute rebuilding criteria between C_i and x ;
  end for
  $b \leftarrow argmax_i(score_i)$;
  test← true if all criteria respected between $score_b$ and acceptRate, false otherwise ;
  $g \leftarrow g + 1$;
end while
return $lambda, mu, sigma$

3.4 Segmentation d'événements ou zones isolés ou atypiques

Les approches de détection de rupture par subdivision hiérarchique dans une série mono- ou multivariée ont démontré leur intérêt lorsque les signaux présentent un régime stationnaire par morceaux (Tsinaslanidis et Kugiumtzis 2014 ; Killick et I. Eckley 2014). Ces ruptures ne sont pas si évidentes dans le cadre des données marines où les événements extrêmes n'imposent pas un changement brutal en moyenne ou variance avec les instants juxtaposés.

La segmentation par classification spectrale bi-classe a montré ses preuves dans la détection de phénomènes communs (exemple de détection de la période productive et non productive phytoplanctonique) ou la détection de bruits ou d'un objet isolé du fond d'une scène (Shi et Malik 2000). Nous avons proposé de revisiter l'approche de clustering bi-récursif de Shi et Malik en proposant un clustering N-divisif basé sur un laplacien L_{NJW} et un critère d'arrêt de qualité de connexité inter- et intra-cluster K. Grassi, **E. Poisson Caillault** et Lefebvre 2019. Ce clustering permet ainsi de hiérarchiser les grandes ruptures dans le signal tout en incluant une phase conjointe de regroupement des sous-séquences obtenues.

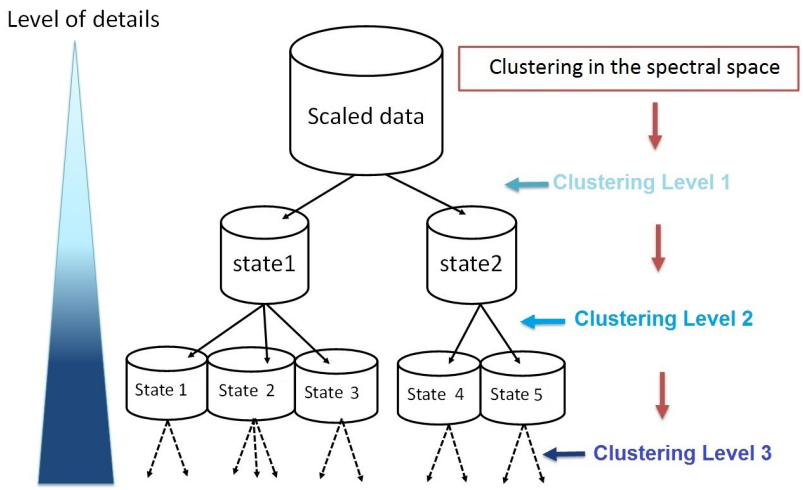


Figure 3.11 – M-SC : Clustering Spectral multi-niveau

Le principe est illustré sur la figure 3.11. Au sommet, l'ensemble de la base est considéré. Le laplacien L_{NJW} est calculé à partir d'une similarité W_{ZP}, dont sont ensuite extraits ses K valeurs propres et vecteurs propres dominants. Dans cet espace spectral, les données sont alors séparées en K groupes par un algorithme K-médoides. Ces groupes représentent les feuilles au niveau suivant qui seront elles-mêmes divisées par le même procédé si la cohésion inter et intra cluster n'est pas suffisante (caractérisée par un seuil d'indice de silhouette à respecter) ou si le nombre de points est inférieur au nombre de voisins considérés dans la matrice ZP (Zelnik-Manor et Perona 2005). L'algorithme nommé M-SC (Multilevel Spectral Clustering) a été testé sur le jeu de données artificiel présenté en début de chapitre composé d'un signal avec trois événements particuliers notés ev1, ev2, ev3. Il a été comparé avec des approches de clustering directs : K-means, NJW-SC et des approches hiérachiques : HC une classification ascendante hiérarchique basée sur la méthode d'aggrégation de Ward appliquée au carré des distances, Hierachical-SC (Sanchez-Garcia et al. 2014) remplaçant l'étape finale de partitionnement spectral-kmeans par un HC, le clustering spectral bi-classe récursif (Bi-SC : Shi et Malik 2000).

À $K = 4$ fixé, aucune méthode testée n'est capable d'isoler ces événements. À partir de $K = 8$, seule l'approche M-SC est capable de les isoler.

Appliqué ensuite au cas MAREL-Carnot, la classification obtenue au niveau 1 correspond au découpage période phytoplanctonique productive/non productive. Le niveau 2 est similaire à la segmentation directe uHMM : schéma non productif/accumulation de nutriments/blooms/fin de blooms. Le niveau suivant permet d'obtenir des événements rares tels

TABLEAU 3.1 – Indicateurs de performance de différents algorithmes de clustering pour segmenter la série artificielle. en gras sont repris les optimaux : ARI - Adjusted Rand Index (ARI), indices de Dunn et Silhouette (Sil.) et scores de précision (Acc.), événements détectés

	ARI	**Dunn**	**Sil.**	**Acc.**	**Evénements détectés**		
Vérité terrain	1	0,008	0,15	-	ev1	ev2	ev3
KM8	0,37	0,010	**0,40**	0,92	35/45	16/38	none
HC8	0,41	**0,022**	0,34	0,90	21/45	none	none
H-SC8	0,42	0,007	0,21	0,89	31/45	none	none
NJW-SC	0,39	0,006	0,26	0,88	36/45	none	none
Bi-SC8	0,33	0,003	0,01	0,88	none	none	none
M-SC8	**0,43**	0,007	0,28	**0,94**	**42/45**	**31/38**	**29/32**

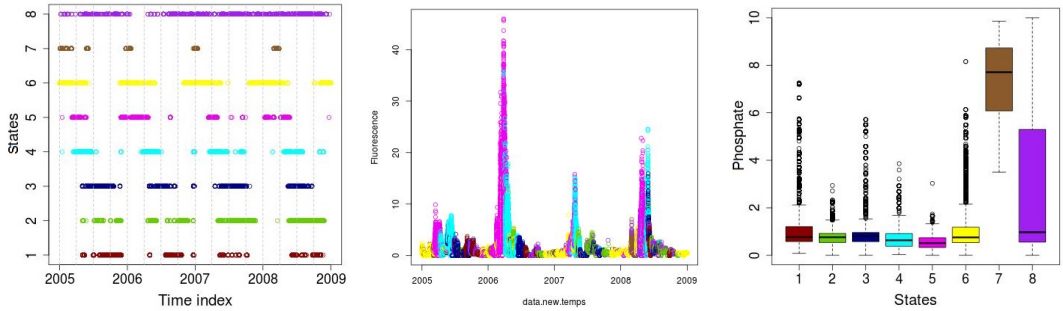


FIGURE 3.12 – Segmentation des données MAREL-Carnot sur la période 2005-2008 obtenu au niveau 3 de l'algorithme MSC).

les désorptions de phosphates, cluster numéro s7 illustré figure 3.12 avec la segmentation obtenu puis la projection colorée sur le signal de fluorescence et la statistique du paramètre Phosphate associée par cluster. Il permet aussi d'identifier le processus impact-réponse des clusters (numéro s4 à s6).

3.5 Conclusions, Perspectives.

La recherche d'événements dans une série spatio-temporelle *sans connaissance a priori* est un sujet ouvert et multiple pour diverses applications. Nous avons proposé un modèle de Markov Caché construit de manière totalement non supervisée et sans hypothèse à partir d'une classification spectrale des données. Cette dernière nous permet de segmenter la série en sous-séquences ayant des propriétés similaires dans les paramètres mesurés et une dynamique équivalente. Plusieurs pistes et travaux ont émergé pour améliorer et/ou contraindre la modélisation obtenue : contrainte temporelle d'assignation entre deux observations mesurées proches, détection d'un nombre d'états minimum et de leurs paramètres dynamiques (probabilité de maintien, probabilité de transitions entre états ou encore d'émission). Ces contraintes de durées pourraient aussi être intégrées dans la formalisation des matrices A et B du HMM (voire π).

D'un point de vue applicatif, le modèle uHMM a été utilisé dans plusieurs contextes : segmentation des événements d'observations longues durées, courtes durées, basse et haute fréquences d'acquisition, bouées fixes ou d'opportunité, campagne de mesures sur bateau.

Bibliographie

E. Poisson Caillault et Lefebvre 2017

E. Poisson Caillault et A. Lefebvre (2017). « Towards Chl-a Bloom Understanding by EM-based Unsupervised Event Detection ». In : *MTS/IEEE Oceans Conference - OCEANS'17 Aberdeen*. DOI : 10.1109/OCEANSE.2017.8084597.

E. Poisson Caillault et Ternynck 2013

E. Poisson Caillault et P. Ternynck (2013). *Package R CRAN uHMM (2016) : détection et modélisation d'événements rares et fréquents dans des séries temporelles multidimensionnelles. Modèle de Markov Caché construit par apprentissage totalement non supervisé. disponible sur https://cran.r-project.org/web/packages/uHMM/index.html*.

G. Wacquet, **E. Caillault Poisson** et al. 2013

G. Wacquet, **E. Caillault Poisson** et al. (2013). « Constrained spectral embedding for K-way data clustering ». In : *Pattern Recognition Letters* 34.9. Impact Factor:1.062, ERA2010: B, h5=, p. 1009–1017. DOI : 10.1016/j.patrec.2013.02.003.

G. Wacquet, **E. Poisson-Caillault** et Hebert 2013

G. Wacquet, **E. Poisson-Caillault** et PA. Hebert (2013). « Semi-supervised K-Way Spectral Clustering with Determination of Number of Clusters ». In : *Computational Intelligence: Revised and Selected Papers of the International Joint Conference, IJCCI 2011, Paris, France, October 24-26, 2011*. Sous la dir. de Kurosh Madani et al. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 317–332. ISBN : 978-3-642-35638-4. DOI : 10.1007/978-3-642-35638-4_21.

K. Grassi, **E. Poisson Caillault** et Lefebvre 2019

K. Grassi, **E. Poisson Caillault** et A. Lefebvre (2019). « Multi-level Spectral Clustering for extreme event characterization ». In : *MTS/IEEE Oceans Conference - OCEANS'19 Marseille*.

K. Rousseeuw, **E. Caillault** et al. 2016

K. Rousseeuw, **E. Caillault** et al. (2016). « Modèle de Markov Caché hybride pour la surveillance de l'environnement marin. » In : *Chapitre d'ouvrage : Mesures à haute résolution dans l'environnement marin côtier*. Edition CNRS ALPHA, p. 164. ISBN : 978-2-271-08592-4.

K. Rousseeuw, **E. Poisson Caillault** et al. 2013

K. Rousseeuw, **E. Poisson Caillault** et al. (2013). « Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling ». In : *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*. h-index:36. Melbourne, Australia, p. 3962–3965. DOI : 10.1109/IGARSS.2013.6723700.

K. Rousseeuw, **E. Poisson Caillault** et al. 2015
K. Rousseeuw, **E. Poisson Caillault** et al. (2015). « Hybrid Hidden Markov Model for Marine Environment Monitoring ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.1. Impact Factor:2.15, p. 204–213. ISSN : 1939-1404. DOI : 10.1109/JSTARS.2014.2341219.

Kévin Rousseeuw 2014
Kévin Rousseeuw (2014). « Modélisation de signaux temporels hautes fréquences multicapteurs à valeurs manquantes : Application à la prédiction des efflorescences phytoplanctoniques dans les rivières et les écosystèmes marins côtiers ». Thèse de doct. supervised by E. Caillault et A. Lebfebvre, dir. D. Hamad, Université du Littoral Côte d'Opale, Boulogne-sur-Mer, France.

Killick et I. Eckley 2014
Killick, Rebecca et Idris Eckley (2014). « changepoint: An R package for changepoint analysis ». In : *Journal of statistical software* 58.3, p. 1–19.

Killick, Fearnhead et I. A. Eckley 2012
Killick, Rebecca, Paul Fearnhead et Idris A Eckley (2012). « Optimal detection of changepoints with a linear computational cost ». In : *Journal of the American Statistical Association* 107.500, p. 1590–1598.

Lefebvre et **E. Poisson Caillault** 2019
Lefebvre, A. et **E. Poisson Caillault** (2019). « High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering ». In : *Marine Ecology Progress Series (IF : 2.276)* 608. DOI : 10.3354/meps12781.

Lefebvre, **E. Poisson-Caillault** et al. 2016
Lefebvre, A., **E. Poisson-Caillault** et al. (2016). « La station instrumentée MAREL Carnot : Retours d'experiences de 10 ans d'observation à haute frequence d'une zone côtière sous influence anthropique. » In : *Chapitre d'ouvrage :Mesures à haute résolution dans l'environnement marin côtier.* Edition CNRS ALPHA, p. 164. ISBN : 978-2-271-08592-4. URL : http://www.cnrseditions.fr/home/7300-mesures-a-haute-resolution-dans-l-environnement-marin-cotier.html.

Prygiel et al. 2018
Prygiel, E. et al. (2018). *Suivi physico-chimique et algal en haute-fréquence du marais d'Isle de Saint-Quentin : Apport de l'interface uHMM pour l'exploitation des données.* Colloque Earth science meeting, 22-26 oct 2018, Lille.

Rabiner 1989
Rabiner, Lawrence R (1989). « A tutorial on hidden Markov models and selected applications in speech recognition ». In : *Proceedings of the IEEE* 77.2, p. 257–286.

Sanchez-Garcia et al. 2014
Sanchez-Garcia, J. et al. (2014). « Hierarchical Spectral Clustering of Power Grids ». In : *IEEE Transactions on Power Systems* 29.5, p. 2229–2237. ISSN : 0885-8950. DOI : 10.1109/TPWRS.2014.2306756.

Shi et Malik 2000
Shi, J. et J. Malik (2000). « Normalized Cuts and Image Segmentation ». In : *IEEE Trans. Pattern Anal. Mach. Intell.* 22.8, p. 888–905. ISSN : 0162-8828. DOI : 10.1109/34.868688. URL : http://dx.doi.org/10.1109/34.868688.

Szekely et Rizzo 2013
Szekely, G.J. et M.L. Rizzo (2013). « Energy statistics: A class of statistics based on distances ». In : *Journal of Statistical Planning and Inference* 143.8, p. 1249–1272. issn : 0378-3758. doi : 10.1016/j.jspi.2013.03.018.

Tsinaslanidis et Kugiumtzis 2014
Tsinaslanidis, Prodromos E. et Dimitris Kugiumtzis (2014). « A prediction scheme using perceptually important points and dynamic time warping ». In : *Expert Systems with Applications* 41.15, p. 6848–6860. issn : 0957-4174. doi : https://doi.org/10.1016/j.eswa.2014.04.028. url : http://www.sciencedirect.com/science/article/pii/S0957417414002516.

Zelnik-Manor et Perona 2005
Zelnik-Manor, Lihi et Pietro Perona (2005). « Self-tuning spectral clustering ». In : *Advances in neural information processing systems*, p. 1601–1608.

Chapitre 4

Conclusions

En écologie numérique, les méthodes de classification sont d'une importance capitale pour synthétiser l'information, comprendre la structure des données et en extraire le maximum d'information à des fins d'amélioration des connaissances et pour émettre des recommandations en matière de gestion de l'environnement. Les connaissances quant au fonctionnement des écosystèmes à de petites échelles de temps sont très souvent incomplètes, ce qui renvoie souvent à préférer les méthodes non supervisées.

Depuis la soutenance de ma thèse en décembre 2015, j'ai changé de thématiques de recherche, de statuts ainsi que de laboratoires. Initialement investie dans l'apprentissage supervisé et discriminant de systèmes à partir de bases de données étiquetées, mes travaux se sont orientés à l'Université du Littoral vers des approches de classification sans ou peu de connaissances *a priori* où l'information de label est inexistante. La classification dite guidée ou encore par paire de contraintes diffère des techniques dites semi-supervisées conventionnelles basées sur un ensemble de données avec un étiquetage partiel.

Projets et collaborations

Ainsi ma thématique scientifique est fortement corrélée aux priorités de l'université du Littoral aux contrées de la science des données, de l'intelligence artificielle et l'environnement marin. Grâce à mon ancrage dans le tissu de partenaires scientifiques régionaux (AEAP - Agence de l'eau Artois Picardie, CEREMA - Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement, IFREMER - Institut Français de Recherche pour l'Exploitation de la Mer, LOG - Laboratoire d'Océanologie et Géosciences) et limitrophes (CEFAS centre de recherche marin en Angleterre, Rijkswaterstaat aux Pays-bas), les travaux de recherche présentés aux chapitres 2 et 3 ont mûri et ont été valorisés dans différents projets repris en annexe dans mon curriculum détaillé. J'ai notamment été responsable et coordinatrice scientifique de l'activité 2, Automatisation des méthodes, du projet INTERREG IVa 2 mers DYMAPHY (2010-2014), Développement d'un système d'observation DYnamique pour la détermination de la qualité des eaux MArines, basé sur l'analyse du PHYtoplancton. De cette collaboration, ont emergé d'un point de vue fondamental l'algorithme cSC de classification spectrale guidée par des contraintes par paire (deux points semblent ou ne doivent pas appartenir au même groupe, G. Wacquet et al. 2013) et l'algorithme uHMM de segmentation et modélisation de séries temporelles en événements caractéristiques ou extrêmes par apprentissage non supervisé d'un Modèle de Markov Caché (K. Rousseeuw et al. 2015). D'un point de vue applicatif, j'ai piloté des transferts vers la communauté scientifique via des logiciels diffusés et des ateliers de formations associés : uHMM, package CRAN et R-ClusTool permettant de partitionner, étiqueter des données de type multiples (séries multidimensionnelles, attributs, images) avec des connaissances expertes de type labels ou paires. Ces ateliers ont engendré diverses communications nationales et internationale énumérées dans l'annexe Bibliographie Personnelle. Suite à une collaboration personnelle forte avec l'institut IFREMER depuis 2009, marquée aussi par une délégation dans le Laboratoire Envi-

ronnement Ressources de Boulogne-sur-Mer, je suis membre officiel du projet JERICO-NEXT ("Towards a Joint European Research Infrastructure Network For Coastal Observatories) et j'interviens dans l'équipe de travail WP3.1 pour développement de plateformes numériques pour l'observation de la diversité phytoplanctonique. Suite à ce projet, j'ai proposé deux actions que je co-pilote dans le prochain INFRAIA JERICO-S3 avec pour perspective l'apprentissage incrémental des anomalies et des événements dans des séries temporelles acquises en flux continu, avec une application dans l'observation de paramètres physico-chimiques acquises sur des transects de bateaux équipés (Lefebvre et **E. Poisson Caillault** 2019). Je reprendrai cette ouverture dans le volet perspective de ce chapitre.

À travers une volonté constante de faire émerger de nouvelles techniques d'extraction de l'information par apprentissage non supervisé et prétraitements avisés des données, j'ai aussi coordonné pour mon laboratoire d'appartenance actuel LISIC l'axe 1 du CPER MARCO - Recherche Marine et Littorale en Côte d'Opale (2010-2014), dédié aux outils d'observation et d'évaluation de l'environnement marin et participé activement au montage de la structure fédérative de recherche SFR MER. J'ai ainsi proposé d'orienter l'évaluation des observations à travers l'angle de la classification, puis de la complétion pour corriger les valeurs aberrantes ou estimer celles manquantes. Les métriques et techniques de déformations élastiques de signaux ont été une des motivations phares afin d'obtenir une explicabilité des processus déployés. De nouveaux algorithmes de complétion de signaux univariés ou multivariés ont été publiés et valorisés dans des dépôts R-CRAN : DTWBI, DTWUMI et FSMUMI détaillés au chapitre 2 de ce document (T.T.H. Phan, **E. Poisson Caillault** et al. 2017 ; T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018). L'apprentissage à partir de comparaison (élastiques ou non) entre signaux a aussi constitué une approche novatrice pour classer des signaux **E. Caillault**, Hebert et Wacquet 2009.

Appliquée à mener des travaux de recherche avec un sens et de les partager, le transfert vers les étudiants est aussi une motivation importante. Les verrous technologiques et les algorithmiques de l'extraction de l'information ont été intégrés aux travers de modules dirigés, encadrements de projets, stages ou thèses mais aussi dans le montage des nouvelles plaquettes de formation. L'ensemble de ces éléments et de mes implications collectives de Recherche et Enseignement est non repris ici pour ne pas interrompre l'exposé des travaux scientifiques menés depuis mon recrutement. Cependant, l'annexe A reprend l'ensemble des contributions pédagogiques, recherche et de responsabilités collectives.

Encadrements doctorals

Avant de présenter mes futures orientations de recherche, je résume ci-après les différentes thèses co-encadrées étayant les méthodes (algorithmes) précitées et leurs applications dans cette conclusion.

— La thèse de Guillaume Wacquet (2007-2010, financement MENRT) a été consacrée au partitionnement de séquences multi-dimensionelles par des approches de classification dans l'espace spectral extrait d'une matrice de similarité. Les similarités élastiques ont permis d'obtenir de bons résultats par rapport aux similarités attributs, sur plusieurs jeux de données UCI et sur une application de partitionnement puis comptage de cellules phytoplanctoniques. Un algorithme de classification spectrale contrainte nommé cSC a été proposé étendant le critère et l'algorithme initial kmeans contraint de Wagstaff (Wagstaff et al. 2001). Cette approche d'ajout de connaissance simple et réduite (données se ressemblant ou totalement différentes) permet d'aider à la fois l'algorithme de partitionnement creusant la matrice de similarité, et la tâche de labellisation par un expert humain, le système lui proposant automatiquement un ensemble de clusters pertinents.

— Dans le cadre de la thèse de Kévin Rousseeuw (financement AEAP-IFREMER, 2011-2014) a été proposé une approche non supervisée, sans utilisation de connaissances environnementales *a priori*, pour construire un système numérique de détection d'états

environnementaux par classification spectrale de données Haute Résolution qui, couplé à une modélisation caché Markovienne, permet de modéliser la dynamique de ces états. Ce système nommé uHMM (et package R) permet de définir des successions d'états environnementaux multicritères (paramètres non corrélés, comme les concentrations en nutriments, la fluorescence, la concentration en oxygène,...), caractéristiques (i) d'états récurrents et/ou extrêmes, (ii) de différentes phases des efflorescences phytoplanctoniques en réponse aux modifications environnementales, (iii) des modifications du milieu en réponse à ces efflorescences.

— La thèse de Thi Thu Hong PHAN (bourse Vietnam, 2015-2018) a été consacrée à la complétion des données manquantes dans des séries multi-dimensions non ou faiblement corrélées. Trois algorithmes fondés sur une hypothèse de cross-corrélation existante dans la série (faible) ont été proposé pour compléter de larges trous dans des séries monovariées ou multivariées. Le principe de ces algorithmes repose sur la construction d'une requête correspondant à la séquence précédent (ou suivant) le trou et la recherche d'un patron similaire. Les deux premiers algorithmes, DTWBI et DTWUMI, sont basés sur une comparaison par métrique élastique, le dernier FSMUMI sur une comparaison via un controleur flou de plusieurs similarités entre deux séquences.

— La thèse CIFRE en cours de Kelly GRASSI (depuis oct. 2017) est dédiée à la détection et prévision d'événements extrêmes à partir d'observations de sources multiples avec des résolutions et paramètres mesurés différents. Une nouvelle architecture divisive de classification spectrale a été proposée et appliquée à la détection d'anomalies et événements extrêmes dans les séries temporelles puis la détection de zones atypiques dans des données géoréférencées. À partir de ces étiquetages automatiques, nous étudierons comment prédire de tels événements à partir de données plus ou moins similaires (satellites et modèles) et sur des sites différents (MAREL-Carnot, MAREL-Iroise, MAREL-MESURHO).

D'autre part, parallèlement à ces thèses, j'ai eu l'occasion de mener des chantiers d'exploration personnelle, dont voici une sélection :

— comparaison des méthodes de complétion, classification et segmentation de séries temporelles ;

— étude et apport des métriques élastiques dans la classification et la complétion de données incomplètes ;

— proposition de méthodes hybrides associant modélisation markovienne et classification spectrale pour aboutir à un modèle uHMM ;

— optimisation du modèle uHMM par intégration de connaissances phénologiques ;

— intégration de connaissances spatio-temporelles dans un apprentissage par classification spectrale non supervisé par une formalisation Must-Link floue.

Perspectives

Les perspectives s'inscrivent dans la continuité de ces travaux antérieurs et dans la définition d'un nouveau projet. Mon objectif est de m'appuyer sur le savoir-faire développé dans les domaines de la classification non supervisé et de la classification supervisée et spécifiquement profonde (très vogue et déjà explorée dans mes travaux de thèses) avec trois chantiers :

— la métrologie intelligente ;

— L'amélioration de l'intégration de connaissances atypiques dans des processus d'aide à la décision ;

— l'apprentissage incrémental de nouveaux phénomènes.

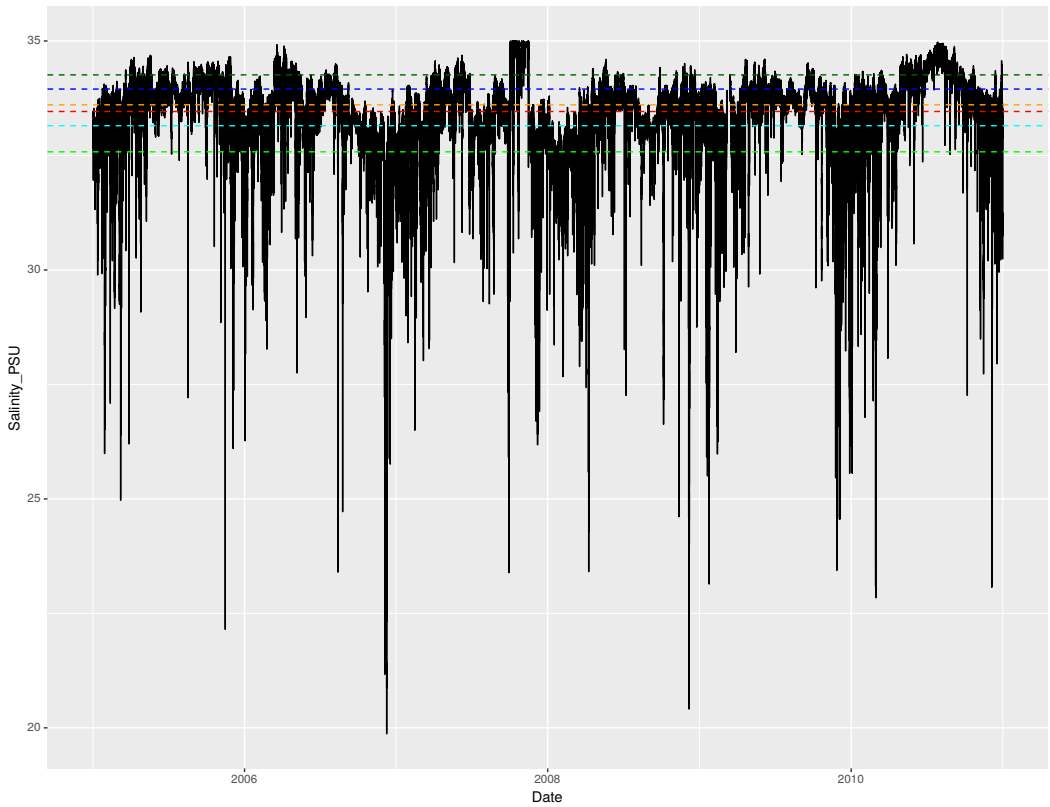


FIGURE 4.1 – Signal de salinité mesurée MAREL-Carnot avec de bas en haut les quantiles 10%, 25%, 50%, 75% et 90% représentés lignes pointillés et la moyenne en rouge.

Pour chacun de ces chantiers, je souhaite intégrer l'aspect gestion de l'incertitude issue à la fois des données, de l'algorithme utilisé et de l'interprétation (étiquetage donné). L'explicabilité des processus et des résultats sont de nos jours un point central pour les potentiels financeurs et certes un sujet en vogue (qui n'a pas cherché à comprendre la machine qu'elle soit de type boîte noire ou simplement complexe).

Métrologie intelligente.

Grâce aux progrès technologiques, l'univers de la métrologie est en plein bouillonnement. "Vérifier puis donner du sens aux données de mesure" était un pilier de l'usine 4.0 toujours d'actualité. Les problèmes de détection des anomalies capteurs présentent des sources de difficultés spécifiques. En effet, pour une information mesurée, se pose la question de la validité de son contenu, de son utilité/pertinence et de l'absence de valeurs manquantes, parmi l'ensemble des acquisitions des capteurs pour répondre à une problématique particulière.

Reprenons le cas du signal de salinité issu des données MAREL-Carnot, illustré à la figure 4.1. Nous ne disposons que d'une mesure certes fréquence acquise toutes les 20 minutes (les autres acquisitions étant faites toutes les deux semaines par les réseaux SOMLIT et/ou REPHY). Les travaux d'étalonnage de capteurs sont donc hors sujet avec de telles différences d'échantillonnage mais aussi de localisation. Il est aisé de détecter une anomalie à la fin de l'année 2007 voire aussi 2010 mais il devient difficile de détecter des erreurs non élémentaires comme certains pics ascendants souvent confondus avec les phénomènes naturels de dessalures dues soit à de fortes pluies soit à des ouvertures de barrage. Nous avons commencé à détecter ces anomalies en les considérant comme des événements extrêmes par le biais de la classification spectrale. Ainsi une anomalie se traduit potentiellement par un label supplémentaire découvert. Récemment dans (MARJUNI, ADJI et FERDIANA 2019), la recherche de défaillance capteur a été initiée directement dans l'espace spectral, recherchant une cohésion autour des

axes. Pour tenter de lever ces sources de difficultés nous envisageons de creuser deux voies. La première concerne la combinaison des approches par rupture et celles de classification. La seconde a pour objet d'apprendre au fil des acquisitions des empreintes courantes valides et ainsi ensuite combiner les approches de comparaison de séquences. Ce volet de Recherche fait l'objet d'un intérêt fort, et nous l'avons proposé dans le projet futur accepté JERICO-S3 en bi-partenariat avec IFREMER, suite du projet actuel JERICO-NEXT.

Intégration de connaissances

La variété des approches et des instruments permet d'enrichir la connaissance d'un processus. Cette complexité "données - algorithmes - interprétation humaine" peut aboutir à différents points de vue - identiques, cohérents, complémentaires, supplémentaires ou à l'inverse totalement orthogonaux. Surviennent alors plusieurs difficultés à résoudre comme l'identification des sources fiables, la co-gestion des informations et des expertises.

Une question simple est comment intégrer des données peu échantillonnées dans des données très discrétisées. Quel sens aura cette donnée ? Sans dégrader un signal par rapport à l'autre, comment exploiter cette nouvelle entrée pouvant être de nature identique (exemple de données satellites, données issus de modèle, données issues de bouées fixes, prélèvements ponctuelles d'un même variable) ou différentes (variables supplémentaires) ?

Chaque type d'information ou connaissance peut être associé à un graphe distinct ou intégré dans le graphe initial. L'intégration des connaissances dans un graphe et la fusion de graphes sont des sujets à fort potentiel renforçant ainsi l'aspect explicabilité individuellement et l'aspect relationnel entre chaque graphe Cai, Zheng et Chang 2018 ; Wang et al. 2017 ; Kazemi et al. 2019 ; Lee et al. 2018. Les graphes d'attention sont notamment intéressants pour intégrer différents points de vue. Nous étudierons notamment cela dans l'intégration de modèles neurologiques, de modèles de réponse aux jeux, des travaux ont été initiés avec la société Orientoi pour le développement d'un agent capable d'apprendre des relations enfants-jeux-métiers.

Apprentissage incrémental

Les modèles générés autrefois à partir d'acquisition à des échelles temporelles ou spatiales passés et les changements brusques ces deux dernières décennies imposent aujourd'hui de repenser nos approches de modélisation et de prédiction mais aussi les hypothèses et conclusions expertes qui ont été faites, publiées.

La combinaison de techniques de classification non supervisée et supervisée (uHMM : modélisation par Modèle de Markov construite à partir d'un clustering spectral) ont déjà apporté une aide intéressante pour l'analyse de séries de mesures et la détection de nouveaux événements rares ou extrêmes (clustering spectral multi-niveau). Cependant elle reste à développer afin d'apporter une interprétation fiable des phénomènes passés et futurs en combinant à la fois des techniques d'Intelligence Artificielle par apprentissage actif incrémental. L'estimation actualisée des phénomènes pourra être orientée par des techniques d'estimation élastique et des techniques de fusion d'information et de décision mais aussi une modélisation des incertitudes tout le long de la chaîne de traitement.

Ces travaux seront intégrés dans les actions liées aux problématiques "Mer et Littoral" (SFR MER) et "Intelligence Artificielle et Optimisation" notamment de l'Alliance A2U (Université Littoral Côte d'Opale, Université d'Artois, et l'université Picardie Jules Vernes).

Bibliographie

Cai, Zheng et Chang 2018

Cai, Hongyun, Vincent W Zheng et Kevin Chen-Chuan Chang (2018). « A comprehensive survey of graph embedding: Problems, techniques, and applications ». In : *IEEE Transactions on Knowledge and Data Engineering* 30.9, p. 1616–1637.

E. Caillault, Hebert et Wacquet 2009

E. Caillault, PA. Hebert et G. Wacquet (2009). « Dissimilarity-Based Classification of Multidimensional Signals by Conjoint Elastic Matching: Application to Phytoplanktonic Species Recognition ». In : *Engineering Applications of Neural Networks - 11th International Conference, EANN 2009, London, UK, August 27-29, 2009. Proceedings*, p. 153–164. doi : 10.1007/978-3-642-03969-0_15.

G. Wacquet et al. 2013

G. Wacquet et al. (2013). « Constrained spectral embedding for K-way data clustering ». In : *Pattern Recognition Letters* 34.9. Impact Factor:1.062, ERA2010: B, h5=, p. 1009–1017. doi : 10.1016/j.patrec.2013.02.003.

K. Rousseeuw et al. 2015

K. Rousseeuw et al. (2015). « Hybrid Hidden Markov Model for Marine Environment Monitoring ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.1. Impact Factor:2.15, p. 204–213. issn : 1939-1404. doi : 10.1109/JSTARS.2014.2341219.

Kazemi et al. 2019

Kazemi, Seyed Mehran et al. (2019). « Relational Representation Learning for Dynamic (Knowledge) Graphs: A Survey ». In : *arXiv preprint arXiv:1905.11485*.

Lee et al. 2018

Lee, John Boaz et al. (2018). « Attention models in graphs: A survey ». In : *arXiv preprint arXiv:1807.07984*.

Lefebvre et **E. Poisson Caillault** 2019

Lefebvre, A. et **E. Poisson Caillault** (2019). « High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering ». In : *Marine Ecology Progress Series (IF : 2.276)* 608. doi : 10.3354/meps12781.

Marjuni, Adji et Ferdiana 2019

Marjuni, Aris, Teguh B. Adji et Ridi Ferdiana (2019). « Unsupervised software defect prediction using median absolute deviation threshold based spectral classifier on signed Laplacian matrix ». In : *Journal of Big Data* 6.1. doi : 10.1186/s40537-019-0250-z. url : https://doi.org/10.1186/s40537-019-0250-z.

T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018
T.T.H. Phan, A. Bigand et **E. Poisson Caillault** (2018). « A New Fuzzy Logic-Based Similarity Measure Applied to Large Gap Imputation for Uncorrelated Multivariate Time Series ». In : *Applied Computational Intelligence and Soft Computing* 2018, 9095683:1–9095683:15. DOI : 10.1155/2018/9095683.

T.T.H. Phan, **E. Poisson Caillault** et al. 2017
T.T.H. Phan, **E. Poisson Caillault** et al. (2017). « Dynamic time warping-based imputation for univariate time series data ». In : *Pattern Recognition Letters (IF: 1.952)*. DOI : 10.1016/j.patrec.2017.08.019.

Wagstaff et al. 2001
Wagstaff, Kiri et al. (2001). « Constrained K-means Clustering with Background Knowledge ». In : *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 577–584.

Wang et al. 2017
Wang, Quan et al. (2017). « Knowledge Graph Embedding: A Survey of Approaches and Applications ». In : *IEEE Transactions on Knowledge and Data Engineering* 29.12, p. 2724–2743. DOI : 10.1109/tkde.2017.2754499. URL : https://doi.org/10.1109/tkde.2017.2754499.

Annexe A

Résumé des activités après thèse

A.1 CV simplifié

Information	née le 19.09.1976 (42 ans) à Carentan (50) divorcée, 2 enfants (12 et 16 ans) mel : emilie.poisson@univ-littoral.fr
Fonctions	2006-... : Maître de Conférences, **ULCO LISIC EA 4491**, Département Informatique Signal 2014-2016 **Délégation partielle IFREMER** au Laboratoire Environnement Ressources à Boulogne-sur-Mer (LER-BL), service complet d'enseignement assuré à l'ULCO.
Diplômes	2006 : Qualifiée CNU 27 et 61 2005 : **Doctorat** en Automatique Informatique Appliquée. Université de Nantes. 2001 : **DEA** Automatique Informatique Appliquée, École Centrale de Nantes 2001 : **Ingénieur** en Systèmes Électroniques et Info. Industrielle, Polytech-Nantes
Enseignement	2006-. . . : Maître de Conférences ULCO 2004-2006 : ATER-61 demi-poste, Polytech-Nantes 2001-2004 : Moniteur-61 CIES Grand Ouest Production et suivi de projets M1, PIC, bibliographiques : 3/an Participation création EILCO et suivi de stagiaires ingénieurs réguliers. Participation montage Licence info Parcours SII, Master TSI, DU ADAM3D (2020)
Recherche	**Mots-clés** : Reconnaissance de formes, classification, modélisation par graphe, modèles de Markov Cachés, algorithme de classification spectrale, apprentissage non guidé, apprentissage semi-supervisé, données multi-dimensionnelles (nD), séries temporelles, complétion, segmentation, supervision de la qualité de l'eau, données marines, phytoplancton.
Projets en cours	- Membre officiel du Projet H2020 Jerico-Next (33 partenaires, 15 pays, 2015-18). - Membre et porteur LISIC axe 1 du CPER MARCO, projet régional, 2016-18.
Encadrements	4 Thèses + 1 débutée en 2019 - divers stages de masters recherche.
Publications	7 Revues, 13 actes de conférences internationales depuis 2006.
PEDR	bénéficiaire de la PEDR de 2017-2020.
Activités collectives	2010-. . . : Membre élue au conseil du laboratoire LISIC 2014-. . . : Membre élue du conseil scientifique du Campus de la Mer 2016-. . . : Membre élue représentant du département EEA puis IS au CGU Calais 2015-. . . : Membre jury VAE Validation Acquis Expérience 2019-. . . : Membre élue représentant Parcours PRREL Licences MSPI du site de Calais 2008-. . . : Resp. Journée Premier Emploi, Resp. site web, com. Master INS3I/TSI 2008-2014 : Directrice des études Master INS3I Membre de divers comités de sélection (2013, 2016, 2019)

A.2 Recherche

Contexte

Ingénieur en Systèmes Électroniques et Informatique Industrielle à Polytech'Nantes en 2001, ma thèse de doctorat de l'Université de Nantes (spécialité Automatique et Informatique Appliquée) a été soutenue en 2005. J'ai proposé plusieurs hybrides neuronaux (TDNN-SDNN) et un Système Hybride Neuro-Markovien pour la Reconnaissance de l'Écriture Manuscrite En-Ligne. Des critères d'apprentissage globaux (par approches conjointes discriminantes et maximum de vraisemblance) ont été proposés et validés sur plusieurs bases internationales.

En 2016, j'ai été recruté en tant que Maître de Conférences 61 ème section à l'Université du Littoral Côte d'Opale (ULCO) dans le laboratoire LASL - Laboratoire d'Analyse des Systèmes du Littoral E.A. 2600 - pour selon le profil : faire de la classification supervisée (SVM, RN, hybride) pour des applications transports. A mon arrivée en janvier 2017, après un congé maternité (de mon deuxième enfant), j'ai intégré une équipe réduite de deux personnes, un professeur et un Maître de conférences dont la nouvelle orientation était sur le plan fondamental vers la classification non supervisée et sur le plan applicatif l' Environnement.
Depuis la fusion du LASL avec le LIL, Laboratoire d'INformatique du Littoral en 2010, j'effectue ma recherche au LISIC (Laboratoire d'Informatique, Signal et Image de la Côte d'Opale) dans l'axe IMAP-Apprentissage et Image, constitué de deux professeurs, deux Maîtres de conférences Habilités à Diriger la Recherche, et huit Maîtres de conférences.

Depuis mon recrutement à l'ULCO, j'ai orienté mes travaux suite à la demande émergente de l'équipe soit la classification spectrale non supervisée et semi-supervisée de signaux multidimensionnels avec suivi temporel ou spatial. Et j'ai inscrit mes travaux, projets dans l'axe prioritaire de l'université l'Environnement et notamment le Campus de la mer avec son CPER associé. Après un investissement bibliographique sur la théorie des graphes-classification spectrale, j'ai réussi à proposer de nouveaux algorithmes de comparaison de signaux et intégrer mes compétences initiales en apprentissage notamment via les hybrides markoviens construits désormais de façon non supervisée.

Bilan résumé

Le tableau suivant reprend l'ensemble des activités de recherche détaillées dans la suite par un bilan quantitatif de mes activités en recherche depuis mon recrutement.

Production scientifique	
Articles en revue JCR	5
Communications internationales avec comité de lecture	13
Communications nationales avec actes	8
Logiciels	5
Autres publications	18
Encadrement	
Thèses	3 soutenues + 2 en cours
Post-doctorat	1
Masters recherche	5

Responsabilité scientifique	
Participation aux projets/contrats	8 (+1 futur) dont 6 Resp.
Organisation de conférences	3
Conférence invitée	1
Fonctions électives Recherche	2

Contributions et production scientifique

Mes travaux s'inscrivent dans les domaines du traitement de signal et de la reconnaissance de formes, et plus particulièrement dans le volet Science des données et apprentissage statistique.

Depuis ma thése, mes travaux sont orientés sur la modélisation et classification de données multidimensionnelles sans connaissance a priori ou guidé par peu de connaissances (interface entre le guidé et la classification semi-supervisée).
D'un point de vue fondamentaux, j'analyse, développe et combine des méthodes de clustering, telles que les algorithmes de classification spectrale et les modèles de Markov Cachés basés sur des attributs définis ou des métriques de comparaison par appariemment élastique.
Depuis ma délégation IFREMER, une grande part de mes travaux concerne les problèmes de complétion et/ou modélisation de séries temporelles incomplètes. De par cette délégation et une collaboration avec l'Agence de l'eau Artois Picardie, j'ai développé un package R-CRAN uHMM afin de valoriser ces approches et être exploitables dans le cadre d'un monitoring de la qualité de l'eau marine, côtière ou fluvial de points fixes ou de transects.

J'oriente actuellement et pour les cinq ans à venir ma Recherche principalement sur la prise en compte des connaissances expertes d'un processus (connaissances des modes Lefebvre et **Poisson-Caillault** 2016 ; **E. Poisson Caillault** et Lefebvre 2017, contraintes temporelles ou spatiales, autres vues, ...) directement dans les algorithmes de modélisation et classification. Je co-encadre actuellement un sujet de thèse CIFRE, tri-partie société WEATHERFORCE-IFREMER-ULCO/LISIC dont l'objet est d'aider à l'interprétation des phénomènes extrêmes. Une nouvelle thèse CIFRE avec la société ORIENTOI a été déposée et acceptée par l'ANRT (Agence Nationale de Recherche et le Technologie). Celle-ci débutera en septembre 2019 et a pour objet d'intégrer les connaissances de modèles psychologiques et de comportements de jeux (en-ligne) dans un outil d'aide à la décision dans l'orientation des élèves (proposition de profil métier).

Classification spectrale guidée par des contraintes de paires. Il est souvent plus aisé de recueillir des comparaisons subjectives d'expert (contraintes : appartient à la même classe ou n'appartient pas à la même classe) que des étiquettes de classes (notamment en biologie le protocole d'identification microscopique est laborieuse voir infaisable). Nous avons proposé un algorithme de classification permettant de générer un sous-espace de projection par optimisation d'un critère de multi-coupe normalisé avec ajustement des coefficients de pénalité dus aux contraintes. Les performances de l'algorithme sont mises en évidence sur différentes bases de données par comparaison à d'autres algorithmes de la littérature (PRL 2013 G. Wacquet, **E. Caillault Poisson** et al. 2013 et NCTA2011P.A. Hébert, **E. Caillault Poisson** et Hamad 2011). Dans (SCI 2013 G. Wacquet, **E. Poisson-Caillault** et Hebert 2013) a été une version de l'algorithme de classification spectrale paramétré totalement de manière automatique, indépendante de l'utilisateur. La partie applicative a été publiée et récompensée dans (STIC 2011 G. Wacquet, Hebert et al. 2011).

Détection d'événements et modélisation Nous avons contribué à la détection et l'extraction intelligente de signatures caractéristiques dynamiques d'épisodes fugace ou répétitif de longueurs variables dans des séries temporelles multi-dimensionnelles. Une modélisation semi-markovienne basée sur un construction non supervisée via une classification de type nuée dynamique a été proposée et validée sur différents jeux de données (une station fixe Carnot MAREL, jeu temporel dans IGARSS 2013 K. Rousseeuw et al. 2013 et un pocket ferrybox, jeu temporel et spatial dans Kevin Rousseeuw et al. 2013). La classification non supervisée permet ainsi à la fois de générer les états d'un Modèle de Markov caché (MMC)et le codebook des symboles d'observations possibles lié à ces états détectés. Ce premier modèle a été généralisé pour des ensembles non convexes avec une approche de classification spectrale. La combinaison hybride classification spectrale-MMC a été publié dans la revue JSTAR et validée pour différentes applications marines, données issues de stations fixes (MAREL2014) ou

mobiles (FBW2013,FBW2014). Un nouvel algorithme de clustering divisif spectral a été proposé pour orienter la segmentation des séries dans les événements extrêmes (OCEANS2019). De nombreux rapports techniques et présentations dans les doctoriales - doctoriales Lille Nord de France (2-7 juin 2013) et les doctoriales de la mer, campus de la mer (10 octobre 2013)- ont été réalisés.

Analyse et amélioration de signaux La qualité d'une donnée est nécessaire pour obtenir un résultat de classification pertinent. Les données ou l'information sont parfois incomplètes, fausses ou périmées. Il est important d'améliorer la qualité du signal d'origine et de choisir des métriques ou attributs de comparaison à la fois généralisantes et discriminants. Nous travaillons particulièrement sur les méthodes de déformation élastique type DTW tant pour compléter des séries temporelles que les comparer (Kévin Rousseeuw 2014-chp2). Dans EANN 2009 **E. Caillault**, Hebert et Wacquet 2009 et STIC 2009 **Caillault** et al. 2009, nous avons construit un nouvelle architecture neuronale qui se base sur un ensemble de comparaison DTW d'une nouvelle donnée par rapport à un set de données étiquetés. Nous avons aussi proposé une nouvelle métrique à la fois qualitative et quantitative pour distinguer deux signaux. Cette approche a été appliqué à différents jeux de données classiques et dans le cas de classification d'espèces phytoplanctoniques à partir de données cytométriques (G. Wacquet, Hebert et al. 2011). Pour ces signaux cytométriques (série temporelle multidimensionnelle à n courbes), nous avons proposé des attributs basé moment discriminants permettant de réduire la complexité de calcul et recherche de l'algorithme précédent. Ces travaux ont été publiés dans ICCE 2016 T.T.H. Phan, **E. Poisson Caillault** et Bigand 2016. Nous avons proposé des méthodes de complétion par recherche d'un motif similaire au phénomène observé par des métriques élastiques T.T.H. Phan, **E. Poisson Caillault**, Lefebvre et al. 2017 ; T.T.H Phan et al. 2017 ou flous T.T.H. Phan, Bigand et **E. Poisson Caillault** 2018 et avec la même méthodologie des agents de prédictions T.T.H. Phan, **E. Poisson-Caillault** et Bigand 2018.

A.3 Mobilité et Collaborations

De septembre 2014 à août 2016, j'ai été reçu en **délégation** partielle à IFREMER LER-Boulogne-sur-mer (service d'enseignement complet assuré).

Fortement engagée dans l'axe prioritaire de l'ULCO sur l'environnement, je me déplace fréquemment dans la région notamment dans le cadre d'une étroite collaboration avec l'IFREMER Boulogne-sur-Mer et le Laboratoire d'Océanologie et Géosciences LOG de Wimereux, WEATHERFORCE et METEOFRANCE à Toulouse (1 semaine en avril 2019) mais aussi au niveau international avec le Pays-Bas (5 semaines de 2010 à 2013 à Middelburg pour l'agence de l'eau hollandaise) et l'Angleterre (2 semaines en 2012 et 2014 au CEFAS, Lowerstoft-Angleterre).

A.4 Encadrement doctoral et scientifique

Encadrement de thèse à venir

— à partir de sept.2019, Siegfried Delannoy
  Co-direction 50% avec A. BIGAND,
  Financement : CIFRE, société ORIENTOI basé à Lille, offrant des outils numériques et ludiques d'aide à l'orientation scolaire et professionnelle.
  Sujet : "Apprentissage dynamique du profil des joueurs et des correspondances joueur - secteur d'activité, joueur - métiers"

Encadrement de thèse en cours.

1. depuis oct.2017, Kelly GRASSI,
   Co-encadrement 50% avec Dr. Alain Lefebvre - IFREMER, sous la direction administrative de A. BIGAND pour l'ULCO,
   Financement : CIFRE, société WEATHERFORCE basée à Toulouse offrant des solutions de prédictions climatologiques à partir d'informations de capteurs existants, détournées de leurs applications initiales (crowd-sensing véhicule ou mobile).
   Sujet : "Caractérisation de la dynamique de la biomasse phytoplanctonique par définition d'états environnementaux multicritères avec apprentissage profond semi-supervisé et classification spectrale à partir de données hautes fréquences"
   Une architecture de clustering spectral divisif multi-coupes a été proposé.
   Production actuelle : 1 conférence et plusieurs communications internationales.

Encadrement de thèses soutenues

2. sept.2015-oct.2018, Thi Thu Hong PHAN,
   Co-direction 50% avec HdR André Bigand,
   Financement : Programme de bouse 911, Ambassade de France - Ministère Vietnamien.
   endt : "Complétion et classification de séries temporelles à partir de méthodes d'appariement élastiques." Plusieurs algorithmes de complétion de série basés sur des similarités élastiques ou floues et une recherche d'un phénomène déjà observé ont été proposés.
   Production associée : 2 revues et 4 conférences internationales et autres.
   Thi Thu Hong est retournée en tant que enseignant chercheur à l'Université d'Agriculture de Hanoï (VNUA -Vietnam National University of Agriculture), et enseigne dans le département Informatique de la faculté des sciences (Faculty of Information technology).

3. déc.2010 - déc.2013 - Kévin ROUSSEEUW,
   Co-encadrement 50% avec Dr. Alain Lefbvre - IFREMER, sous la direction administrative de Pr. Denis Hamad,
   Financement : 50% IFREMER - 50 % AEAP
   Sujet : "Modélisation de signaux temporels haute fréquence multi-capteurs à valeur manquantes. Application à la prédiction des efflorescences phytoplanctoniques dans les rivières et les écosystèmes marins côtiers". Les travaux portent sur la construction automatisée et non supervisée d'un modèle de Markov Caché pour la détection et l'extraction intelligente de signatures caractéristiques dynamiques des efflorescences phytoplanctoniques (temporellement).
   Production associée : 1 revue, 4 conférences internationales et autres.
   Kévin a travaillé (en CDI) dans la société Intelligent Video Software (IVS, Lille) où il a mis en oeuvre des algorithmes de classification pour de la supervision. Aujourd'hui et depuis 2018, il est co-fondateur de la société ORIENTOI.

4. oct.2009-dec.2011 : Guillaume WACQUET,
   Co-encadrement 50% avec Pr. Denis Hamad
   Financement : 100% MESR
   Sujet : "Classification spectrale semi-supervisée. Application à la surveillance de l'écosystème marin". Un nouvel algorithme de classification spectrale semi-supervisée a été proposé à partir de contraintes d'association ou de refus d'association de paires de point.
   Production associée : 1 revue, 3 conférences internationales et autres.
   Après avoir été en contrat d'Ingénieur de Recherche à l'Université de Mons en Belgique, Guillaume est actuellement Chercheur post-doctorant au Laboratoire d'Océanologie et de Géosciences - UMR CNRs 8187 LOG. Il a par ailleurs été qualifié en 61ème section

CNU.

Encadrement Post-doc et ingénieur

1. avril.2019-fev.2020 - Ingénieur d'études, Pierre TALON.
   Encadrement 80% avec Pierre-Alexandre Hébert.
   Financement : Projet CPER MARCO.
   Déploiement d'une interface de clustering non supervisé, guidé par contraintes à supervisé pour des données multi-sources (attributs, courbes, images) sous R.

2. 2018 - Ingénieur de Recherche, Camille DEZECACHE (Dr.)
   Encadrement 100%, pilotage de l'axe 1 pour le LISIC
   Financement : Projet CPER MARCO.
   Déploiement des outils de complétion sous R.

3. 2016-17 - Ingénieur d'études, Paul TERNYNCK puis Quentin MARSON.
   Encadrement 100%
   Financement : Convention AEAP - ULCO/LISIC. Déploiement d'une interface de segmentation de séries par combinaison de modélisation markovienne et clustering spectral sous R.

Encadrements de projets et stages de Master Recherche.

1. Mémoire M2 de P. Chatelain, "Suivi de clusters de phytoplanctons", réalisé au LISIC, avril-sept 2019. Encadrement 80%, financement Projet CPER MARCO.
2. Mémoire M2 de A. Rizik, "How to Insert Temporal Information on Spectral Clustering Algorithm." Stage de Master 2 Research STIP- Signal, Telecoms., Image and Speech) de L'université Libanaise -Faculty of Sciences I (Hadath), réalisé au LISIC, avril-août 2016. Encadrement 80%.
3. Mémoire M2 de C. Herbez, "Traitement de données manquantes ou aberrantes, Imputation multiple. Applications aux données Marel." Projet Master INS3I 2ère année (100h), ULCO, LISIC équipe EIA, 2012. *Il a soutenu sa thèse de doctorat en 2016 sous la direction du Pr. Éric Ramat au LISIC-équipe Osmose, Optimisation Simulation MOdeliSation Evolutionnaire.* Encadrement 100%.
   Production associée : Rédaction d'un rapport technique et présentation orale lors d'un séminaire lié au projet DYMAPHY (**E. Poisson Caillault** 2013).
4. Mémoire M2 de K. Vermast. Identification du scripteur à partir d'un document numérisé. Projet Master INS3I 2ère année (100h), ULCO, LISIC équipe EIA, 2011. Encadrement 100%.
5. Mémoire M2. X. Yang. Concentration de cellules phytoplantoniques à partir d'acquisition issu d'un fluoroprobre. Projet Master INS3I 2ère année (100h), ULCO, LISIC équipe EIA, 2011. Encadrement 100%.
6. Mémoire M1. X. Yang. Segmentation et classification d'images de cellules phytoplanctoniques issus d'une caméra couplée à un cytomètre. Stage Master 1ère année, ULCO, LISIC équipe EIA, 2011. Encadrement 100%.
7. Mémoire M1. M. Delannoy, J. Bonnard. Prototypage intelligent en vue d'une identification d'un auteur. Projet Master INS3I 1ère année (100h), ULCO, LISIC équipe EIA, 2010. Encadrement 100%.
8. Mémoire M2. G. Mollet. Prédiction de données manquantes et aberrantes dans des séries temporelles dépendantes. Projet Master INS3I 1ère année (100h), ULCO, LISIC équipe EIA, 2010. Encadrement 100%.

9. Mémoire DEA. S. Péchard, "Psychovisual controf of spatial mode prediction in H264", Stage de DEA Automatique Informatique Appliquée, École Centrale de Nantes, 2004. Encadrement 33%.
   Production associée : P. LeCallet, C. Viard-Gaudin, S. Pechard and **E. Caillault**. "No reference and reduced Reference video quality metrics for end to end QoS monitoring". in IEICE Transactions on Communications, special issue on Multimedia QoS Evaluation and Management Technologies. Volume E85. February 2006.
10. Mémoire DEA. F. Alleau. "Time-delay neural networks workshop and features extration on videos". Stage de DEA Automatique Informatique Appliquée, École Centrale de Nantes, 2003. Encadrement 33%.
    Production associée : F. Alleau, **E. Poisson**, C. Viard-Gaudin et P. Le Callet, "TDNN with Masked Inputs", in Proc. of the 4th International Conference on Information, Communications & Signal Processing and IEEE Pacific-Rim Conference on Multimedia (ICICS-PCM 2003), Nanyang Technological University, Singapour, Decembre 2003.

A.5 Projets, Contrats.

Depuis mon arrivée à l'ULCO 2007, j'ai participé activement à plusieurs projets de recherche de visibilité locale, régionale et nationale : Classpec 2008, BQR Jeune Chercheur 2008, BQR PhytoClass 2008-2010. Depuis 2012, j'ai participé à des projets de portée frontalière européenne, régionale et internationale. Impliquée dans le projet actuel JERICO-Next, j'ai participé au nouvel appel à projet et suis co-responsable de deux Workpackages du projet JERICO 3 (en cours d'évaluation).

1. **Participation scientifique** du projet Claspec, jan.2007-dec.2008 : Classification spectrale pour la segmentation d'images couleur et de signaux audio.
   Montant du projet : 12kfinancé par le Grayshym Porteur : Pr. D. Hamad
   Partenaires : EMD-École des Mines de Douai, INRETS-Institut national de recherche sur les transports et leur sécurité devenu IFSTARR, LAGIS-Laboratoire d'Automatique, Génie Informatique et Signal devenu CRYSTal, HEI-École des hautes études d'ingénieur et ULCO-Calais. *Implication personnelle :*
   - Participation aux dix groupes de lectures
   - co-orgnaisatrice du Workshop (journée associée à LFA'08) ClasSpec08 : Spectral adn Fuzzy Clustering Techniques : Application to signal and image segmentation. Lens-Wednesday, October 15th 2008. Invitation de 4 conférenciers.
   - Étude comparative des formalismes utilisés en spectral clustering, présentation poster et rapports.
2. **Responsable scientifique** de l'activité 2 Classification, détection automatique du **Projet INTERREG VIa** DYMAPHY, 2008-2013 : Développement d'un système d'observation DYnamique pour la détermination de la qualité des eaux MArines, basé sur l'analyse du PHYtoplancton.
   Projet financé le FEDER (426 k).
   Projet Interreg IV A, 2 Mers, 2010-2013.
   Partenaires : Laboratoire d'Océanologie et Géosciences (LOG UMR CNRS 8187, Wimereux 62), Université du littoral Côte d'Opale – LISIC, Université Lille 1, IFREMER Boulogne, CEFAS Lowerstoft Angleterre, Rijkswaterstaat Middelburg Hollande. 27 personnes.
   Coordinateurs du projet : L. F. Artigas (LOG-ULCO) et D. Hamad (LISIC-ULCO).
   Implication personnelle :
   - Implication dans la gestion et l'analyse statistique des bases de données
   - Rédaction des rapports finaux de l'activité 2

— Développement d'algorithmes de classification non supervisée guidée par l'expert.

— Développement des deux interfaces R pour la cytométrie et l'analyse des phytoplanctons par fluorométrie spectrale.

— Mise en place d'un atelier sur les outils de classification.

— Coordination de la tâche 2, déplacements réguliers et collaboration. Présentations orales et écrites des avancées.

Ce projet commence a porté ses fruits, notamment en terme de publication (ASLO 2015 Artigas et al. 2015, PhycoTox2015 Bonato et al. 2015), nous continuons à travailler ensemble dans le cadre du CPER MARCO et le projet JERICO-Next.

3. **Participation scientifique** dans le **projet H2O2O** Jerico-Next notamment WP3 et WP4, membre officiel comme chercheur associé à IFREMER LER Boulogne-sur-mer, 2015-2020. Réseau Européen Joint d'Infrastructures de Recherche pour les Observatoires Côtiers – Nouvelle Expertise.
Implication personnelle :

   — Implication dans la gestion des bases de données

   — Proposition d'outils et algorithmes d'analyse de séries temporelles issus de stations fixes et mobiles.

   — Participation, chairman de session du workshop international sur les avancées en cours dans l'application de techniques (semi-) automatisées pour le suivi de la dynamique phytoplanctonique en eaux côtières et marines, 31 mai au 2 juin 2016 à Wimereux Lefebvre, Grosjean et al. 2016.

4. **Coordinatrice d'axe** 1 pour le LISIC du CPER MARCO (4 axes), 2016-2020, projet labellisé par le pôle AQUIMER : Recherches marines et littorales en Côte d'Opale : des milieux aux ressources, aux usages et à la qualité des produits aquatiques.
Partenaires : IFREMER LER Boulogne, ULCO-LISIC Calais, Agence de l'Eau Artois Picardie, Laboratoire d'Océanologie et Géosciences (LOG UMR CNRS). 11 personnes pour l'axe 1
Porteur du CPER : F. Schmitt (LOG), porteur axe 1 LISIC : E. Caillault (ULCO-LISIC).
Projet contrat de plan Etat-Région, montant de la partie que je pilote : 55kEuros

   — Je pilote le volet Modélisation et suivi temporel de la dynamique des efflorescences phytoplanctoniques dans l'axe 1.

   — Développement de nouveaux algorithmes d'extraction d'attributs discriminants de cytogrammes, de clustering et de détection d'événements intermittents dans des séries multi-capteurs.

Contrats.

1. **Participation scientifique** à convention PhytoClass, 2008-2009 : Détection de phytoplanctons connus et de nouvelles espèces par classification semi supervisée à partir de données cytométriques.
Projet en collaboration avec le Laboratoire d'Océanologie et de Géosciences (LOG UMR CNRS 8187).
Porteurs : Pr. Denis Hamad.
Achat d'une caméra dédiée au cytomètre en flux.
Implication personnelle :
- Co-encadrement d'un étudiant de Master.
- Développement d'un algorithme de comparaison de cytogrammes(EANN2009, STIC 2009).

2. **Porteuse du BQR Jeune Chercheur ULCO**, 2007-2008 : Identification du scripteur à partir d'un échantillon d'écriture en-ligne par approche globale.
   Financement : 3k, achat d'une tablette de collecte de signatures et texte et financement d'une inscription pour une compétition.
   Implication personnelle (totale) : - Développement d'un algorithme d'identification, participation à la compétition SigComp09 ICDAR 2009.
   - Encadrement d'un étudiant de master, stage et projets.
3. **Co-coordinatrice du Projet Marel Carnot**, 2012-2014 : Apprentissage de signaux temporels. Application à la modélisation de la dynamique des efflorescences phytoplanctoniques dans un écosystème côtier anthropisé.
   Projet financé IFREMER/AEAP/ULCO (426 k).
   Partenaires : IFREMER LER Boulogne, ULCO-LISIC Calais, Agence de l'Eau Artois Picardie, Laboratoire d'Océanologie et Géosciences (LOG UMR CNRS). 11 personnes.
   Coordinateurs du projet : E. Caillault (ULCO-LISIC) et A. Lefebvre (IFREMER)
   Financement de la thèse de Kevin Rousseeuw, fonctionnement et conférences.
   4 soumissions ont été présentées au colloque Marel Carnot 2014.
4. **Coordinatrice** de la Convention ULCO-LISIC/Agence de l'eau Artois Picardie 2015-2017 : Valorisation des outils suite à la thèse de Kevin Rousseeuw.
   Financement : AEAP/ULCO (53 k), Financement d'un ingénieur d'étude et de formation-terrain.
   Partenaires : ULCO-LISIC Calais, Agence de l'Eau Artois Picardie. 9 personnes
   Coordinateurs du projet : E. Caillault (ULCO-LISIC) et J. Prygiel (Agence de l'eau).
   Implication personnelle :
   - Encadrement d'un ingénieur de recherche - Contribution au développement d'un package CRAN R et d'une interface de détection et modélisation des événements usuels et extrêmes dans les séries temporelles d'eau douce ou marine.

A.6 Responsabilités collectives et Rayonnement scientifique

Membre de conseil

— Depuis 2010 : Membre élue au conseil du laboratoire LISIC.
  Participation à la rédaction du nouveau réglement intérieur.

— Depuis 2014 : Membre élue du conseil scientifique du GIS Campus de la Mer.

— 2017-2018 : Membre suppléante élue du laboratoire pour représenter le LISIC dans la création et montage de la SFR MER.

Organisation de colloque

— **Co-organisatrice et comité de programme** du Workshop (journée associée à LFA'08) ClasSpec08 : Spectral adn Fuzzy Clustering Techniques : Application to signal and image segmentation. Lens-Wednesday, October 15th 2008. Invitation de 4 conférenciers.

— **Organisatrice** d'une rencontre-atelier "Data analysis Meeting" dans le cadre du projet DYMAPHYsur les outils développés : Data analysis Meeting, 20-22 mars 2013, LISIC-Calais. (17 personnes)
  Implication personnelle :
  - session Tutoriels et présentations des outils ;
  - session Analyses des campagnes, comparaison des classifications manuelles et automatiques ;
  - Animation d'une Matinée Conférenciers sur la parcimonie et les problèmes inverses.
  Invités associés : Partenaires DYMAPHY, Chercheurs aux MIO : Dr M. Thyssen et G.

Gregori. (Mediterranean Institute of Oceanography, Marseille), un ingénieur de Cytobuoy (Netherlands),équipe SYVIP du LISIC : Dr G. Roussel, Dr G. Delmaire, Dr M. Puigt.

— **Tutoriel et conférencière invitée** : RESOMAR 2013 Atelier Pelagos RESOMAR 2013 du 4 au 6 décembre à Wimereux, France .
Organisatrice de la session pratique : identification du phytoplancton à partir de l'analyse de données fluorométriques et cytométriques par des techniques de classification semi- automatisées
- 3 présentations et organisation de deux sessions parallèles avec un public large (50 personnes : chercheurs, biologistes, physiciens). Mise en pratique des plateformes R citées ci-dessus sur des échantillons marins ou de culture passés la veille dans les instruments cytomètre en flux et fluoromètres spectraux.

— **Chairman** de la demi-journée doctorale du Campus de la Mer, 20 octobre 2017, Boulogne-sur-Mer.

— **Co-organisatrice** de la journée annuelle scientifique du GIS Campus de la Mer : La Méthodologie au service de la mer qui aura lieu le 3 avril 2017 à Boulogne-sur-mer.
Co-organisateur : E. Caillault (LISIC) et A. Lefebvre et B. Ernande (IFREMER).

Participation aux GDR

Je suis membre de trois groupes de recherche : le GRCE - Groupe de Recherche en Communication Ecrite, le GDR ISIS - Groupe de Recherche Information, Signal, Images et ViSion et le GRAISHYM -Groupe d'Intérêt Scientifique de Recherche en Automatisation Intégrée et Systèmes Homme-Machine.

— **Suivi et Participation aux Journées GRAISHYM**
Participation aux journées REPAR 2015 et 2016.
Présentation orale à la journée Image vision et RDF,2 juin 2011 : Comparaison d'images à partir de similarités des projections mojettes » (JIV2011).

— **Participation au GDR PhytoCox**, 2015 et 2016. Deux présentations orales.

J'ai également présenté mes travaux de recherche dans quatre séminaires au sein du laboratoire devant différents partenaires tels que INNOCOLD, CAP GEMINI, Agence de l'eau Artois Picardie, IFREMER Boulogne, Brest et Caen et trois à l'extérieur (LOG, IFREMER, Cytobuoy Netherlands).

Diffusion logicielle et valorisation

J'ai managé et participé au développement de plusieurs packages R de classification et modélisation de séries temporelles ou données environnementales. Ces packages et fonctions mises à disposition de la communauté permettent ainsi de diffuser nos travaux à des non expert en programmation ou en classification.

1. Package R FCMUMI (2018) : Imputation of Time Series Based on Fuzzy Logic, imputation de séries temporelles univariées et multivariées par recherche d'une récurrence du phénomène observé selon une fusion floue de similarité.
2. Package R DTWUMI (2018) : Imputation of Multivariate Time Series Based on Dynamic Time Warping, imputation de séries temporelles multivariées et peu corrélées par recherche d'une récurrence du phénomène observé selon une métrique élastique.
3. Package R DTWBI (2018) : Imputation of Time Series Based on Dynamic Time Warping, imputation de séries temporelles univariées par recherche d'une récurrence du phénomène observé selon une métrique élastique.

4. Package R CRAN uHMM (2016) : détection et modélisation d'événements rares et fréquents dans des séries temporelles multidimensionnelles. Modèle de Markov Caché construit par apprentissage totalement non supervisé. disponible sur https://cran.r-project.org/web/packages/uHMM/index.html. Développeurs : E. Poisson, P. Ternynck et K. Rousseeuw.
5. Package R-FCM (2013, refonte en cours) : classification supervisée de données cytométriques, classification intéractive (implémentation des méthodes proposées de classification spectrale semi- supervisée), interface de comparaison de classification manuelles et automatiques. Développeurs : PA. Hébert, E. Poisson Caillault, G. Wacquet.
6. Package R-Fluorométrie (2013) : validation d'empreintes fluorométriques, identification d'empreintes, estimation des concentrations d'espèces. Développeurs :E. Poisson Caillault, PA. Hébert.
   Ces deux packages sont disponibles via le site web Dymaphy (http://www.dymaphy.eu/) et ont été distribués aux participants lors des conférences RESOMAR et le colloque de clôture de DYMAPHY sur clé usb.
7. participation aux compétitions internationales en 2009 Blankers et al. 2009 ICDAR : Signature Verification Competition et en 2010 : Forensic Verification Competition 4NSigComp2010 Liwicki et al. 2010.

Participation dans les comités de programmes ou review

— Reviewer régulière de
- Revue PRL, Pattern Recognition Letters
- Revue J-STARS, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing
- Revue Journal of Electronic Imaging

— Reviewer occasionnelle pour les conférences et journaux suivants :
- Revue JARS, Journal of Applied Remote Sensing
- Revue ESWA - Expert Systems With Applications
- Revue IJMLC, International Journal of Machine Learning and Computing.
- ESANN, European Symposium on Artificial Neural Networks.
- Conférence CORESA, COmpression et REprésentation des Signaux Audiovisuels.

— Comité de programme et sélection pour
- Conférence Réunion des Sciences de la Terre, 2018.
- Conférénce DYMAPHY 2013.
- Colloque MAREL 2014.
- Journée CLASPEC 2008.

— Évaluations de projet
- Commission post-doctorale IFREMER 2019

— Participation au Comité de suivi des thèses
- Monica MICHELRODRIGUEZ, "Étude de la dépendance et photoacclimatation chromatique des microalgues. Conséquences sur les estimations de production primaire en mers côtières. LOG CNRS. 2018

A.7 Enseignement depuis 2006

A.7.1 Responsabilités pédagogiques

- Directrice des études de 2008-2014 du Master INS3I - Ingénierie Numérique, Signal Image et Informatique Industrielle - qui suite à la nouvelle habilitation en 2014 devenu master TSI, Traitement du signal et des Images :
  - Suivi pédagogique des étudiants : étude des dossiers d'inscriptions, suvi des étudiants, conseils d'insertion à l'emploi ou l'orientation doctorale, jury d'informations sur le master, etc.
  - Tâche de coordination : emploi du temps, absences, réservation de salles, lien avec la scolarité, mise en place et suivi du respect des modalités de contrôle et connaissance, etc.
  - Tâche de communication : Création des affiches et diaporama de la formation, Mise en place de la première Journée Premier Emploi et participations aux JPE suivantes de 2014 à ce jour, Mise en place du site web, facebook, et d'un Partenariat Pictanovo.
  - Rédaction de la nouvelle habilitation de la formation TSI avec le Professeur Dominique Schneider.
  - Coordinateur des stages et projets du Master INSI. (sans décharge horaire).
- Depuis 2014, membre élue à la commission Validation Acquis d'Expérience (VAE). (sans décharge horaire).
  - étude des dossiers, environ 30 par an.
  - participation au jury de validation, 4 minimum par an.
- Depuis 2014, membre de l'EILCO, école d'ingénieurs du Littoral Côte d'Opale.
  - Participation au montage du dossier CTI de la formation Ingénieur et du cycle préparatoire intégré Informatique. Définition de fiches (des contenus) des modules du Parcours Informatique Industrielle.
  - Montage de nouveaux cours et TPs : Systèmes à événements discrets, Systèmes temps-réel & Java.
  - Suivi des projets PIC, projets bibliographiques et stages-ingénieur.
- Membre des comités de sélection IUT Béthune 2016, EILCO 2013 (CNU 61) et 2019 (CNU 61), ULCO 2019 (CNU 27).
- Depuis 2012, mise en place de partenariat avec des anciens diplômés et industriels pour présenter en L1 les métiers de l'informatique industrielle dans le cadre des modules PPP, en Master les ateliers Métiers sous forme de conférence et d'échanges avec les étudiants ou projets pratiques.
- Membre du jury de Licence 1 Informatique et anciennement de la licence SPI E4I.
- Depuis 2016, membre élue du département EEA Génie Mécanique au e Conseil Consultatif du CGU Calais-Longuenesse.
- Depuis 2019, membre élue Représentant PRREL (Parcours Région Réussite des Etudes Longues) pour les licences MSPI du site de Calais.

A.7.2 Mission d'orientation et d'information

- Participation aux forums d'information sur l'enseignement supérieur :
  - 2007 - 2014 : Participation aux journées Portes Ouvertes de l'ULCO. Mise en place d'une salle dédiée à la formation Licence SPI puis informatique et Master TSI.

  - 2013, 2014 : Salon de l'étudiant Calais, espace Gambetta. Animation des sessions : Que faire avec un bac scientifique ? La place des femmes en sciences.
  - 2012, 2014 : Journée organisée par le lycée Branly Boulogne-sur-Mer et ouverts à tous les lycéens de 1ère S du département.
  - Lycée de l'Europe Dunkerque, présentation de la licence informatique et promotion de la femme en sciences.
  - Lycée Berthelot : Promotion des masters et des orientations possibles après un BTS.
- Réception et conseil tout au long de l'année des étudiants pour leur orientation
  - Démonstration aux étudiants de L1 des métiers de la Recherche et des applications associées, notamment dans le cadre de mes travaux démonstration des différentes interfaces logicielles développées pour le suivi de la qualité de la mer.
  - Accueil régulier de collégiens de quatrième et troisième dans le cadre de leur stage découverte pour promouvoir le métier d'enseignant chercheur en informatique appliquée.
  - Correction des curriculum vitae avant la JPE- Journée Premier Emploi, et au fil de l'eau analyse et correction des dossiers de candidature des étudiants aux concours et écoles (filières Histoire, Informatique).

A.7.3 Résumé des enseignements

Les tableaux suivants reprennent l'ensemble des activités d'enseignements depuis la thèse. L'université et ses antennes ou départements étant sur 4 sites, j'enseigne principalement sur Calais mais aussi Dunkerque, Boulogne-sur-mer et Longuenesse. J'ai la chance de pouvoir couvrir un large public de la licence 1 au doctorat avec différents cadres : universitaire, IUT, école d'ingénieurs.

Tableau A.1 – Liste des enseignements de 2017 à 2019.

eq.TD	**Type**	**Formation**	**Intitulé de l'enseignement**
2018-2019 : mcf Université du Littoral Côte d'Opale			
45 h	CM+TD+TP	Licence 1 Informatique	Architecture des ordinateurs
9h	TD	Licence 3 Histoire	PPP unité 3
17h	TD	Master 1 Histoire	PPP unité 4
12.08h	CM+TD+TP	Master 1 Sciences de la Mer	Programmation sur R
30	CM+TD+TP	Master 1 TSI	Systèmes à événements discrets
30	CM+TD+TP	Master 1 TSI	Temps réel
46	CM+TD+TP	INGénieur 3 GI1 et INFO1 EILCO	Architecture des Ordinateurs
5	TD	INGénieur 4 GI2 EILCO	ETUDE TECHNIQUE
2	TD	INGénieur 4 EIL INFO2 EILCO	Bureau d'études
51	CM+TD	INGénieur EIL INFO2 EILCO	Temps réel
10	TD	INGénieur 5 EIL INFO3 EILCO	Bureau d'étude industriel
2	TD	INGénieur 5 EIL INFO3 EILCO	Projets
6	TD	INGénieur 5 EIL INFO3 EILCO	Stages
2017-2018 : mcf Université du Littoral Côte d'Opale			
2	TD 1	Référentiel Orientation	JPO - Journées Portes Ouvertes
36	CM+TD+TP	Licence 1 Informatique	Architecture des ordinateurs
10	TD	Licence 3 Histoire BOULOGNE	PPP unité 3
17	TD	Master 1 Histoire	PPP Unité 4
12.08	CM+TD+TP	Master 1 Sciences de la Mer	Programmation sur R
30	CM+TD+TP	Master 1 TSI	Systèmes à événements discrets
30	CM+TD+TP	Master 1 TSI	Temps réel
46	CM+TD+TP	ING 3 EIL GI1 et INFO1 EILCO	Architecture des Ordinateurs
5	TD	ING 4 EIL GI2 EILCO	ETUDE TECHNIQUE
51	CM+TD	ING 4 EIL INFO 2 EILCO	Temps réel
10	TD	ING 5 EIL INFO3 EILCO	Bureau d'étude industriel

Tableau A.2 – Liste des enseignements de 2014 à 2017).

eq.TD	Type	Formation	Intitulé de l'enseignement
2018-2019 : mcf Université du Littoral Côte d'Opale			
2016-2017 : mcf Université du Littoral Côte d'Opale			
24 h	CM+TP	Licence 1 Informatique	Architecture des ordinateurs
12 h	TP	Licence 1 Informatique Dk	Architecture des ordinateurs
9 h	TD	Licence 3 Histoire Boulogne	PPP unité 3
17 h	TD	Master 1 Histoire	PPP
12.08 h	CM+TD+TP	Master 1 Sciences de la Mer	Programmation sur R
30 h	CM+TD+TP	Master 1 TSI	Systèmes à événements discrets
30 h	CM+TD+TP	Master 1 TSI	Temps réel
42 h	TP	DUT 2ère année Informatique	M4102C Programmation Répartie
36 h	CM+TD+TP	EIL INFO 2	Temps réel
2015-2016 : *en délégation IFREMER*			
2 h	TD	Référentiel Orientation	JPO
24 h	CMTP	Licence 1 Informatique	Architecture des ordinateurs
12 h	TP	Licence 1 Informatique Dk	Architecture des ordinateurs
9 h	TD	Licence 3 Histoire Boulogne	PPP unité 3
17 h	TD	Master 1 Histoire	PPP
12.08 h	CM+TD+TP	Master 1 Sciences de la Mer	Programmation sur R
30 h	CM+TD+TP	Master 1 TSI	Systèmes à événements discrets
30 h	CM+TD+TP	Master 1 TSI	Temps réel
2 h	TD	Master 2 TSI	Projet
7 h	TP	DUT 1ère année Informatique	Bases de la programmation orientée objet
12 h	TD	EIL INFO 2	Stages
36 h	CM+TD+TP	EIL INFO 2	Temps réel
1.5 h	TD	EIL GI3	Projets
6 h	TD	EIL INFO 3	Stages
2014-2015 : *en délégation IFREMER*			
2 h	TD	Référentiel Orientation	SUAIO-JPO
1 h	TD	Référentiel Orientation	SUAIO-Salons
33 h	CM+TD+TP	Licence 1 Informatique	Architecture des ordinateurs
6 h	TD	EIL INFO 2	Stage
30 h	CM+TD+TP	EIL INFO 2	Temps réel
22.5 h	CM	Master 1 INS3I -	Méthodologie UMLRT
1 h	TD	Master 1 INS3I -	Projet
37.5 h	CM+TD	Master 1 INS3I -	Systèmes à événements discrets
22.5 h	CM	Master 1 INS3I -	Systèmes d'exploitations Temps Réel
18.75 h	CM+TD	Master 2 INS3I -	Classification de données et Représentation
20 h	TD	Master 2 INS3I -	Outils libres
2 h	TD	Master 2 INS3I -	Projet
15 h	TD	Master 2 INS3I	Réseaux de neurones et app. statistique
2013-2014 : mcf, Université du Littoral Côte d'Opale			
18 h	TD	Référentiel Formation	Resp. année M1 INS3I
18 h	TD	Référentiel Formation	Resp. année M2 INS3I
0.5 h	TD	Référentiel Orientation	Actions Orientation
13 h	CM+TD	Licence 1 MSPI	E4I
3 h	TD	Licence 3 SPI E4I	Stages
16 h	TD	EIL INFO 2	Stages
15 h	CM	EIL INFO 2	Temps réel
21 h	CM+TD+TP	Master 1 Informatique	Systèmes temps réel
22.5 h	CM	Master 1 INS3I	Méthodologie UMLRT
4 h	TD	Master 1 INS3I	Projet
37.5 h	CM+TD	Master 1 INS3I	Systèmes d'exploitation Temps Réel
18.75 h	CM+TD	Master 1 INS3I	Classification de données et représentation
15 h	TD	Master 2 INS3I	Réseaux de neurones et app. statistique
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesures

Tableau A.3 – Liste des enseignements de sept. 2010 à août 2013

eq.TD	Type	Formation	Intitulé de l'enseignement
2012-2013 : mcf, Université du Littoral Côte d'Opale			
18 h	TD	Référentiel Formation	Resp. année M1 INS3I
18 h	TD	Référentiel Formation	Resp. année M2 INS3I
25 h	CM+TD	Licence 1 MSPI	E4I
37 h	CM+TD+TP	Licence 2 SPI E4I	Automatique et Automatisme
30 h	CM+TD+TP	EIL INFO 1	Systèmes à événements discrets
8 h	TD	EIL INFO 2	Stages
24 h	CM+TD	EIL INFO 2	Temps réel
29.17 h	CMTP	Master 1 INS3I	Méthodologie UMLRT
4 h	TD	Master 1 INS3I	Projets
37.5 h	CM+TD	Master 1 INS3I	Systèmes d'exploitation temps réel
18.75 h	CM+TD	Master 2 INS3I	Classification de données et représentation
20 h	TD	Master 2 INS3I	outils libres
6 h	TD	Master 2 INS3I	Projets
18.75 h	CM+TD	Master 2 INS3I	Réseaux de neurones et app. statistique
3 h	TD	Master 2 INS3I	Stages
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesures
2011-2012 : mcf, Université du Littoral Côte d'Opale			
25 h	CM+TD	Licence 1 MSPI	E4I
21 h	CM	Licence 2 SPI E4I	Automatique et Automatisme
2 h	TD	Licence 2 SPI E4I	Colles Automatisme et automatique
32.5 h	CM+TD+TP	EIL INFO 1	Systèmes à événements discrets
29.17 h	CMTP	Master 1 INS3I	Méthodologie UMLRT
4 h	TD	Master 1 INS3I	Projet
4 h	TD	Master 1 INS3I	Stage
37.5 h	CM+TD	Master 1 INS3I	Systèmes d'exploitations temps réel
18.75 h	CM+TD	Master 2 INS3I	Classification de données et représentation
20 h	TD	Master 2 INS3I	outils libres
6 h	TD	Master 2 INS3I	projet
18.75 h	CM+TD	Master 2 INS3I	Réseaux de neurones et app. statistique
2 h	TD	Master 2 INS3I	stage
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesures
2010-2011 : mcf, Université du Littoral Côte d'Opale			
2.25 h	TD	Licence 1 MSPI	Dispositions Pédagogique Paritaires S1
50 h	CM+TD	Licence 1 MSPI	E4I
7 h	TD	Licence 1 MSPI	Interrogations Orales E4I
6.67 h	TP	Licence 1 MSPI	SI E4I
5 h	TP	Licence 1 MSPI - Dk	SCIENCES DE L'INGENIEUR-II
21 h	CM	Licence 2 SPI E4I	Automatique et Automatisme
2 h	TD	Licence 2 SPI E4I	Colles - Automatisme et automatique
1 h	TD	Licence 2 SPI E4I	DPP - Automatique et automatique
44.17 h	CM+TD+TP	Master 1 INS3I	Méthodologie UMLRT
2 h	TD	Master 1 INS3I	Projet
2 h	TD	Master 1 INS3I	Stage
37.5 h	CM+TD	Master 1 INS3I	Systèmes d'exploitations temps réel
22.5 h	CM	Master 2 INS3I	Form. et impl. des systèmes distribués
20 h	TD	Master 2 INS3I	Outils libres
4 h	TD	Master 2 INS3I	Projet
18.75 h	CM+TD	Master 2 INS3I	Réseaux de neurones et app. statistique
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesures

Tableau A.4 – Liste des enseignements de jan. 2007 à août 2010

eq.TD	Type	Formation	Intitulé de l'enseignement
2009-2010 : mcf, Université du Littoral Côte d'Opale			
25 h	CM+TD	Licence 1 MSPI	EEA
7.5 h	TD	Licence 1 MSPI	SPI EEA
29 h	CM+TD	Licence 2 ST EEA	Automatique et Automatisme
1 h	TD	Licence 2 ST EEA	Colles Automatique et Automatisme
44.17 h	CM+TD+TP	Master 1 INS3I	Méthodologie UMLRT
10 h	TD	Master 1 INS3I	Projet
2 h	TD	Master 1 INS3I	Stage
37.5 h	CM+TD	Master 1 INS3I	Systèmes d'exploitations
22.5 h	CM	Master 2 INS3I	Form. et impl. des systèmes distribués
20 h	TD	Master 2 INS3I	outils libres
18.75 h	CM+TD	Master 2 INS3I	Réseaux de neurones et app. statistique
2 h	TD	Master 2 INS3I	Stage
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesures
25 h	CM+TD	Licence 1 MSPI	EEA
6 h	TD	Licence 1 MSPI	SPI EEA
29 h	CM+TD	Licence 2 ST EEA	Automatique et Automatisme
1.25 h	TD	Licence 2 ST EEA	Dispositions Paritaires Pédagogiques
2008-2009 : mcf, Université du Littoral Côte d'Opale			
44.17 h	CM+TD+TP	Master 1 INS3I	Méthodologie UMLRT
6 h	TD	Master 1 INS3I	Projet
4 h	TD	Master 1 INS3I	Stage
37.5 h	CM+TD	Master 1 INS3I	Systèmes d'exploitation temps réel
6 h	TD	Master 2 INS3I	Projet
18.75 h	CM+TD	Master 2 INS3I	Réseaux de neurones et app. statistique
4 h	TD	Master 2 INS3I	Stage
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesures
2007-2008 : mcf, Université du Littoral Côte d'Opale			
25 h	CM+TD	Licence 1 MSPI	EEA
1 h	TD	SUAIO	Action Lycée
44.17 h	CM+TD+TP	Master 1 INS3I	Méthodologie UMLRT
4 h	TD	Master 1 INS3I	Projet
8 h	TD	Master 1 INS3I	Stage
37.5 h	CM+TD	Master 1 INS3I	systèmes d'exploitation temps réel
30 h	CM+TD	Master 2 ISIDIS	Fouille de données
4 h	TD	Master 2 INS3I	Projet
18.75 h	CM+TD	Master 2 INS3I	Réseaux de neurones et app. statistique
6 h	TD	Master 2 INS3I	Stage
18.75 h	CM+TD	Master 2 INS3I	Systèmes embarqués et télémesure
2006-2007 : mcf, *en congé Maternité sept-dec. 2016*			
15 h	TD	Licence 3 EEA	Bureau d'études
44.17 h	CM+TD+TP	Master 1 INS3I	Identification des systèmes
11 h	TD	Master 1 INS3I	Projet/stage
37.5 h	TD+TP	Master 1 INS3I	Systèmes d'exploitation temps
4 h	TD	Master 2 INS3I	Stage

Annexe B

Bibliographie Personnelle depuis 2006

B.1 Publications dans des journaux JCR

B.2 Chapitres d'ouvrages

B.3 Publications dans des conférences internationales avec comité de sélection (full paper)

B.4 Publications dans des conférences avec comité de sélection (sur résumé) et Communication depuis 2017

B.5 Logiciels

Annexe B- Bibliographie personnelle depuis 2006

Sont soulignés les doctorants encadrés, en gras mon nom ortographié (Caillault, Poisson Caillault ou Caillault Poisson, ...). Seules les communications depuis 2017 sont reportées, les années antérieures sont visibles la toile.

Articles dans revues internationales à comité de lecture - JCR

[1] A. Lefebvre and **É. Poisson Caillault**. "High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering". In: *Marine Ecology Progress Series (IF : 2.276)* 608 (Jan. 2019). DOI: 10.3354/meps12781.

[2] T.T.H. Phan, A. Bigand, and **É. Poisson Caillault**. "A New Fuzzy Logic-Based Similarity Measure Applied to Large Gap Imputation for Uncorrelated Multivariate Time Series". In: *Applied Computational Intelligence and Soft Computing* 2018 (2018), 9095683:1–9095683:15. DOI: 10.1155/2018/9095683.

[3] T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre, and A. Bigand. "Dynamic time warping-based imputation for univariate time series data". In: *Pattern Recognition Letters (IF: 1.952)* (Aug. 2017). DOI: 10.1016/j.patrec.2017.08.019.

[4] K. Rousseeuw, **É. Poisson Caillault**, A. Lefebvre, and D. Hamad. "Hybrid Hidden Markov Model for Marine Environment Monitoring". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.1 (Jan. 2015). Impact Factor:2.15, pp. 204–213. ISSN: 1939-1404. DOI: 10.1109/JSTARS.2014.2341219.

[5] G. Wacquet, **E. Caillault Poisson**, D. Hamad, and P. Hébert. "Constrained spectral embedding for K-way data clustering". In: *Pattern Recognition Letters* 34.9 (2013). Impact Factor:1.062, ERA2010: B, h5=, pp. 1009–1017. DOI: 10.1016/j.patrec.2013.02.003.

[6] **E. Caillault** and C. Viard Gaudin. "Mixed Discriminant Training of Hybrid and ANN/HMM systems for online handwritten word recognition". In: *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)* 21.1 (2007), pp. 117–134. DOI: 10.1142/S0218001407005338.

[7] P. LeCallet, C. Viard-Gaudin, S. Pechard, and **E. Caillault**. "No Reference and Reduced Reference Video Quality Metrics for End to End QoS Monitoring". In: *IEICE Transactions on Communications* E89-B.2 (Feb. 2006), pp. 289–296. DOI: 10.1093/ietcom/e89-b.2.289.

Chapitres d'ouvrages

[8] K. Rousseeuw, **É. Caillault**, A. Lefebvre, and D. Hamad. "Modèle de Markov Caché hybride pour la surveillance de l'environnement marin." In: *Chapitre d'ouvrage : Mesures à haute résolution dans l'environnement marin côtier.* Édition CNRS ALPHA, 2016, p. 164. ISBN: 978-2-271-08592-4. URL: http://www.cnrseditions.fr/home/7300-mesures-a-haute-resolution-dans-l-environnement-marin-cotier.html.

[9] A. Lefebvre, **É. Poisson-Caillault**, K. Rousseeuw, D. Hamad, D. Soudant, A. Soudant, F. Gohin, and M. Repecaud. "La station instrumentée MAREL Carnot : Retours d'expériences de 10 ans d'observation à haute fréquence d'une zone côtière sous influence anthropique." In: *Chapitre d'ouvrage :Mesures à haute résolution dans l'environnement marin côtier.* Édition CNRS ALPHA, 2016, p. 164. ISBN: 978-2-271-08592-4. URL: http://www.cnrseditions.fr/home/7300-mesures-a-haute-resolution-dans-l-environnement-marin-cotier.html.

[10] G. Wacquet, **E. Poisson-Caillault**, and PA. Hébert. "Semi-supervised K-Way Spectral Clustering with Determination of Number of Clusters". In: *Computational Intelligence: Revised and Selected Papers of the International Joint Conference, IJCCI 2011, Paris, France, October 24-26, 2011.* Ed. by Kurosh Madani, António Dourado, Agostinho Rosa, and Joaquim Filipe. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 317–332. ISBN: 978-3-642-35638-4. DOI: 10.1007/978-3-642-35638-4_21.

Conférences internationales avec comité de lecture (full paper)

[11] K. Grassi, **E. Poisson Caillault**, and A. Lefebvre. "Multi-level Spectral Clustering for extreme event characterization". In: *MTS/IEEE Oceans Conference - OCEANS'19 Marseille.* 2019.

[12] T.T.H Phan, **E. Poisson Caillault**, and A. Bigand. "eDTWBI: effective imputation method for univariate time series". In: *International Conference on Computer Science, Applied Mathematics and Applications - ICCSAMA'2019, Hanoi, Vietnam.* 2019.

[13] T.T.H. Phan, **É. Poisson Caillault**, and A. Bigand. "Comparative Study on Univariate Forecasting Methods for Meteorological Time Series". In: *2018 26th European Signal Processing Conference (EUSIPCO).* Sept. 2018, pp. 2380–2384. DOI: 10.23919/EUSIPCO.2018.8553576.

[14] **E. Poisson Caillault** and A. Lefebvre. "Towards Chl-a Bloom Understanding by EM-based Unsupervised Event Detection". In: *MTS/IEEE Oceans Conference - OCEANS'17 Aberdeen.* 2017. DOI: 10.1109/OCEANSE.2017.8084597.

[15] T.T.H. Phan, **E. Poisson Caillault**, A. Bigand, and A. Lefebvre. "DTW-Approach for uncorrelated multivariate time series imputation". In: *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP).* Sept. 2017, pp. 1–6. DOI: 10.1109/MLSP.2017.8168165.

[16] T.T.H. Phan, **E. Poisson Caillault**, A. Lefebvre, and A. Bigand. "Which DTW Method Applied to Marine Univariate Time Series Imputation". In: *MTS/IEEE Oceans Conference - OCEANS'17 Aberdeen.* 2017. DOI: 10.1109/OCEANSE.2017.8084598.

[17] T.T.H. Phan, **E. Poisson Caillault**, and A. Bigand. "Comparative study on supervised learning methods for identifying phytoplankton species". In: *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE).* July 2016, pp. 283–288.

[18] K. Rousseeuw, **E. Poisson Caillault**, A. Lefebvre, and D. Hamad. "Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling". In: *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*. h-index:36. Melbourne, Australia, July 2013, pp. 3962–3965. DOI: 10.1109/IGARSS.2013.6723700.

[19] G. Wacquet, PA. Hébert, **E. Caillault Poisson**, and D. Hamad. "Semi-supervised K-way Spectral Clustering using Pairwise Constraints". In: *NCTA 2011 - Proceedings of the International Conference on Neural Computation Theory and Applications [part of the International Joint Conference on Computational Intelligence IJCCI 2011], Paris, France, 24-26 October, 2011*. 2011, pp. 72–81.

[20] G. Wacquet, PA. Hébert, **E. Caillault**, and D. Hamad. "Classification semi-supervisée pour l'identification de cellules phytoplanctoniques". In: *STIC et Environnement, Colloque Sciences et Techniques de l'Information et de la Communication Pour l'Environnement*. 2011.

[21] **E. Caillault**, PA. Hébert, and G. Wacquet. "Dissimilarity-Based Classification of Multidimensional Signals by Conjoint Elastic Matching: Application to Phytoplanktonic Species Recognition". In: *Engineering Applications of Neural Networks - 11th International Conference, EANN 2009, London, UK, August 27-29, 2009. Proceedings*. 2009, pp. 153–164. DOI: 10.1007/978-3-642-03969-0_15.

Conférences nationales ou internationales sur résumé, congrès, GDR

[22] D. Devreker, K. Grassi, **E. Poisson-caillault**, A. Lefebvre, and K. Grassi. *Présentation des outils numériques développés pour l'exploitation des données haute-fréquence FerryBox*. CGFS workshop, 25 avril 2019. Boulogne-sur-mer, France., 2019.

[23] **E. Poisson Caillault**, K. Grassi, and A. Lefebvre. *Spectral clustering multi-level for ST event learning*. Artificial Intelligence Workshop. Université d'artois, CRIL, 29 août 2019, 2019.

[24] **E. Poisson-Caillault**, K. Grassi, and A. Lefebvre. *Machine Learning et Dynamique Phytoplanctonique*. Congrés ULCO, Avenir Littoral. 13 Mars, Dunkerque, France, 2019.

[25] K. Grassi, **E. Poisson Caillault**, A. Bigand, and A. Lefebvre. *Détection de zone(s) atypique(s) dans des données géoréférencées par classification spectrale*. Journées MARCO. Nausicaa, Boulogne-sur-mer, Oct. 2019.

[26] K. Grassi, **E. Poisson-Caillault**, and A. Lefebvre. *Machine Learning et observatoire marin*. Séminaire Méteo-France. Avril 9, 2019. Toulouse, France, 2019.

[27] K. Grassi, **E. Poisson-Caillaut**, and A. Lefebvre. *Extreme Event detection from multivariate data time series. Application to marine observation.* Journée IA atmosphère océan climat, Fev 5-7, 2019- Rennes, France. 2019.

[28] A. Lefebvre, K. Grassi, D. Devreker, and **E. Poisson-caillault**. *Phytoplancton, hydrologie et Haute Fréquence : développements numériques*. Séminaire Ifremer : Fondation Tara Océan, 5 juin 2019. Maison de l'Océan, Paris., 2019.

[29] A. Lefebvre, K. Grassi, and **E. Poisson-Caillault**. *Identification of spatial Hydro-biological structures by spectral clustering. Towards implementation of machine learning for Ferry Box data processing. FerryBox Workshop*. Genoa, Italie., 2019.

[30] **E. Poisson Caillault**, A. Bigand, and A. Lefebvre. *Time series segmentation and learning.* Séminaire IA, LISIC, Université du Littoral Côte d'Opale, 18 Juin 2018. 2018.

[31] Grassi, K., **Poisson Caillault, E.**, A. Bigand, and A. Lefebvre. *Multi-level spectral clustering for extreme event discovery.* Séminaire IA, LISIC, Université du Littoral Côte d'Opale, 18 Juin 2018. 2018.

[32] K. Grassi, T.T.H. Phan, **E. Poisson-Caillault**, A. Bigand, D. Devreker, and A. Lefebvre. *Results from measurements in the Eastern English Channel MAREL Carnot-station.* MIO, Marseilles., 2018.

[33] A. Lefebvre, K. Grassi, T.T.H. Phan, D. Devreker, A. Bigand, and **E. Poisson-Caillault**. *Automated tools for analyzing outputs of automated sensors: High frequency Data.* Third JERICO-NEXT Workshop on Phytoplankton Automated Observation. March 19-21, 2018. M.I.O., Marseille, France., 2018.

[34] PHAN, H., **Caillault, E.**, A. Lefebvre, and A. Bigand. *Multivariate times series completion by unsupervised and supervised way.* Séminaire IA, LISIC, Université du Littoral Côte d'Opale, 18 Juin 2018. 2018.

[35] E. **Poisson-Caillault**, K. Grassi, T.T.H. Phan, C. Dezecache, J. Prygiel, and A. Lefebvre. *DTWBI and uHMM R-packages for multivariate time series preprocessing and interpretation.* Colloque Earth science meeting, 22-26 oct 2018, Lille. 2018.

[36] E. Prygiel, E. **Poisson-Caillault**, C. Halkett, E. Chermette, and Prygiel J. *Suivi physico-chimique et algal en haute-fréquence du marais d'Isle de Saint-Quentin : Apport de l'interface uHMM pour l'exploitation des données.* Colloque Earth science meeting, 22-26 oct 2018, Lille. 2018.

[37] G. Wacquet, A. Louchart, C. Blondel, P.-A. Hébert, **Poisson-Caillault, E.**, F. Gomez, Lefebvre A., Ph. Grosjean, and L.F. Artigas. *Combination of "machine learning" methodologies and automated data acquisition systems for phytoplankton detection and classification.* AG JERICO NEXT, Galway. Sept. 2018.

[38] A. Lefebvre A., D. Devreker, K. Grassi, and **E. Poisson-Caillault**. *Analyse de tendance et classification spectrale couplée à un modèle de Markov caché.* Colloque EVOLECO : EVOlution à Long terme des Ecosystèmes COtiers : Vers une mise en évidence des forçages et des processus associés, 5-7 décembre 2017, Bordeaux. 2017.

[39] Karlson B., Anglès S., Artigas L.F., Brosnahan M.L., Colas F., Creach C., Reinhoud De Blok R., Deneudt K., Eikrem W.and Hällfors H.and Gregori G.and Kielosto S.and Kuosa H., Laakso L., Lefebvre A., Lehtinen S., Louchart A., Oja J., Rijkeboer M., **Poisson-Caillault E.**, Jukka Seppälä J., Suikkanen S., Tamminen T., Thyssen M., Tyberghein Land Wacquet G., and Ylöstalo P. *Flow cytometry and imaging in flow methods facilitate automated observations and monitoring of algal blooms and phytoplankton abundance and diversity in automated platforms.* FerryBox workshop on Color Fantasyand

Norwegian Institute for Water Researchand Osloand 17-19 October 2017. 2017.

[40] Artigas L. F., Bonato S., Claquin P., Créach V., de Blok R., Deneudt K., Dugenne M., Grégori G., Grosjean P., Hamad D., Hébert P.-A., Houliez E., Karlson B., Kromkamp J., Lahbib S., Lefebvre A., Lizon F., Louchart A., Ove Möllerand K., Petersen W., **Poisson-Caillault E.**, Revilla M., Rijkeboer M., Rutten T., Tyberghein L., Thyssen M., Seppälä J., Stemmann L., Veen A., Vywerman W., Wacquet G., and Wollschläger J. *Automated approaches for studying phytoplankton dynamics in coastal marine waters: single-particle vs. bulk optical sensors within the JERICO-Next H2020 network. CytoBuoy International Workshop. Woerdenand Netherlandsand March 27-29.* 2017.

[41] Artigas L.F., Bonato S., Claquin P., Créach V, de Blok R., Deneudt K., Dugenne M., Grégori G., Grosjeanand P.and Hamad D., Hébert P.-A., Houliez E., Karlson B., Kromkamp J., Lahbib S.and Lefebvre A., Lizon F., Louchart A., Ove Möllerand K.and Petersen W., **Poisson-Caillault E.**, Revilla M., Rijkeboer M., Rutten T.and Tyberghein L., Thyssen M., Seppälä J., Stemmann L., Veen A., Wacquet G.and Wollschläger J., and Vywermanand W. *Automated characterisation of phytoplankton dynamics in coastal marine waters: the DYMAPHY project and the JERICO-Next H2020 network.* International Council for the Exploration of the Sea – Working Group on Phytoplankton Microbial Ecology. Reykjavikand Icelandand March 28-30. 2017.

Rapports techniques

[42] A. Lefebvre and **E. Poisson-Caillault**. *MAREL Carnot : Rapport n 12 : Bilan d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan 2017. Ifremer/RST.LER.BL/18.05, 24 pages.* Tech. rep. 2018.

Logiciels - Packages

[43] C. Dezecache, T.T.H. Phan, and **E. Poisson-Caillault**. *DTWBI: Imputation of Time Series Based on Dynamic Time Warping.* 2018. URL: `https://cran.r-project.org/web/packages/DTWBI/index.html`.

[44] C. Dezecache, T.T.H. Phan, and **E. Poisson-Caillault**. *DTWUMI: Imputation of Multivariate Time Series Based on Dynamic Time Warping.* 2018. URL: `https://cran.r-project.org/web/packages/DTWUMI/index.html`.

[45] T.T.H. Phan, A. Bigand, and **E. Poisson-Caillault**. *FSMUMI: Imputation of Time Series Based on Fuzzy Logic.* 2018. URL: `https://cran.r-project.org/web/packages/FSMUMI/index.html`.

[46] **E. Poisson Caillault**, PA. Hébert, and G. Wacquet. *Package R-FCM (2013, refonte en cours -> RClusTool) : classification supervisée de données cytométriques, classification intéractive (implémentation des méthodes proposées de classification spectrale semi- supervisée), interface de comparaison de classification manuelles et automatiques.* 2013. URL: `www.dymaphy.eu`.

[47] **E. Poisson Caillault**, PA. Hébert, and G. Wacquet. *Package R-Fluorométrie : validation d'empreintes fluorométriques, identification d'empreintes, estimation des concentrations d'espèces. Développeurs :E. Poisson Caillault, PA. Hébert. disponible sur le site* `www.dymaphy.eu`. 2013.

[48] **E. Poisson Caillault** and P. Ternynck. *Package R CRAN uHMM (2016): détection et modélisation d'événements rares et fréquents dans des séries temporelles multidimensionnelles. Modèle de Markov Caché construit par apprentissage totalement non supervisé. disponible sur* `https://cran.r-project.org/web/packages/uHMM/index.html`. 2013.

Annexe C

Liste des acronymes

BG moyenne géométrique des écarts– ou *Geometric Mean Bias*–. 22

cSC Constrained Spectral Clustering - classification spectrale contrainte. 44

EOV Variables Océaniques Essentielles – ou Essential Ocean Variable. 5

FA2 Facteur 2. 22
FB biais fractionnel – ou *Fractional Bias*–. 22
FS fraction des écarts-types – ou *Fractional Standard deviation*–. 22
FSD fraction des écarts-types – ou *Fractional Standard Deviation*–. 22

HMM Hidden Markov Model. 11, 13

IFREMER Institut Français de Recherche pour l'Exploitation de la Mer. 4

MAE moyenne des écarts absolus – ou *Mean Absolute Error*–. 22
MB biais moyen – ou *Mean Bias*–. 22
ME moyenne des écarts – ou *Mean Error*–. 22

NMAE moyenne des écarts absolus normalisée– ou *Normalized Mean Absolute Error*–. 22
NMB biais moyen normalisé– ou *Normalized Mean Bias*–. 22

R^2 coefficient de détermination. 22
RMSE erreur quadratique moyenne des écarts – ou *Root Mean Square Error*–. 22

VG variance géométrique– ou *Geometric Mean Variance*–. 22