Big data et science des données pour le suivi des ressources naturelles

Jannaï Tokotoko

Année 2022

Ecole Doctorale Du Pacifique

Thèse

Présentée devant
l'Université de la Nouvelle Calédonie
par

Jannaï TOKOTOKO

pour obtenir le diplôme de

Docteur spécialité Informatique

Big data et science des données pour le suivi des ressources naturelles

Soutenue publiquement le 12 Décembre 2022, devant le jury composé de :

Président : Babau, Jean-Philippe

Examinateurs : Laugier, Thierry, Fournier-Viger, Philippe

Co-directeurs

Nazha SELMAOUI-FOLCHER, Maître de Conférences, HDR, ISEA / UNC
Hugues LEMONNIER, Chercheur, ENTROPIE / IFREMER

Sommaire

Liste des figures

Liste des tables

Chapitre 1

Introduction

La multiplication des données collectées ces dernières années, les avancées récentes en matière d'observation à distance (p.ex. imagerie satellitaire très haute résolution), et les nouvelles générations de capteurs et d'appareils connectés peu coûteux, laissent entrevoir un grand nombre de possibilités en matière de suivi et de gestion des ressources (p.ex. suivi de l'évolution des cultures et des espaces protégés). Toutefois, elles soulèvent aussi un grand nombre de défis, notamment en matière d'analyse des données, de par la complexité des phénomènes observés et des données collectées (massives, hétérogènes, distribuées, imprécises, bruitées). Les outils scientifiques et technologiques actuels (méthodes statistiques classiques etc.) sont peu adaptés pour gérer et exploiter ces données complexes et massives. Face à ces défis, la science des données (data science) vise à apporter des solutions (méthodes, algorithmes, outils logiciels, etc). Elle permet le croisement de données hétérogènes et l'identification de relations cachées (sans hypothèses a priori) entre ces données.

La science des données se développe en forte interaction entre les thématiciens producteurs des données et les informaticiens. Une synergie entre les scientifiques de différentes disciplines est nécessaire pour relever ces défis. La thèse proposée s'inscrit dans ce contexte. Elle propose une stratégie pour analyser des données complexes générées pour le suivi des ressources naturelles, et de nouvelles méthodes théoriques (quand cela est nécessaire) et de nouveaux outils informatiques pour traiter et analyser ces données. Cette thèse se focalisera plus particulièrement sur le problème du suivi et de la gestion des bassins aquacoles de crevette en Nouvelle-Calédonie. Elle s'appuiera pour cela sur l'expertise en science des données des chercheurs en informatique du laboratoire ISEA (Institut des Sciences Exactes et Appliquées) de l'Université de la Nouvelle-Calédonie, sur la connaissance des problématiques aquacoles des chercheurs de l'équipe LEAD-NC (Lagons, Ecosystèmes et Aquaculture Durable) de l'IFREMER, et sur les acteurs de la filière aquacole (Groupement des Fermes Aquacoles). Elle exploitera pour cela, différentes sources de données disponibles telles que les données d'élevage et les données qualité de la société de commercialisation la SOPAC. L'analyse croisée de ces données vise à mettre en évidence des typologies de réussite des élevages, en vue d'améliorer la qualité des produits et les résultats zootechniques. Ces résultats

permettront la construction d'outils d'aide à la décision à l'échelle des fermes et à l'échelle de la filière.

Les principaux types de données sur lesquels cette thèse se focalise sont des données multi-variées, statiques, temporelles, multi-échelles, décrivant un ensemble de paramètres environnementaux (température, salinité, ,etc.), biologiques et zootechniques. Ces données concernent aussi des indicateurs (statiques et temporels) sur la qualité des produits destinés à l'alimentation humaine. L'acquisition de ces données est réalisée dans un contexte industriel. Ces données sont relevées et analysées dans le but principal d'assurer la rentabilité biologique de la filière les produisant, et d'améliorer la qualité des produits. Le processus de production implique l'utilisation de protocoles et d'outils de suivi de qualité afin d'assurer un produit sans danger pour la santé. Des données sur la qualité des produits sont de ce fait relevées en cours et en fin de cycle de production. D'autres données sur la qualité du milieu d'élevage intéressent les producteurs afin d'avoir une meilleure compréhension du lien entre les variations de paramètres physico-chimiques, comme le pH, la survie et la croissance de l'espèce élevée.

Les questionnements que se posent les acteurs du domaine aquacole en Nouvelle-Calédonie et auxquelles cette thèse tentera de répondre par l'utilisation de méthodes en science de données sont, par exemple :

- Quels sont les paramètres spatio temporels qui améliorent la qualité de la production de la filière ?
- Comment identifier des facteurs (locaux, globaux, environnementiaux, ...) qui influent sur la production de crevettes ?
- Comment identifier des indicateurs de productivités ?
- Comment caractériser un bon élevage ?
- Comment améliorer la rentabilité de la production calédonienne de crevette ?

1.1 Les systèmes de production

L'agriculture désigne, de manière générale, les différents systèmes de production liés à la culture de plantes et à l'élevage d'animaux. L'aquaculture produit des organismes aquatiques, poissons, mollusques, crustacés et plantes aquatiques inclus. Un système de production est défini comme étant un regroupement de systèmes d'exploitation individuels disposant à peu près d'un même niveau de ressources, pratiquant les mêmes modes de production, bénéficiant des mêmes sources de subsistance, et assujettis aux mêmes contraintes, pour lesquels des stratégies et interventions de développement similaires peuvent être élaborées.

Au cours des quatre dernières décennies du 20ème siècle, la population du monde a presque doublé, comptant 8 milliards d'habitants en 2022, elle devrait approcher 10 milliards en 2050.

L'agriculture et l'aquaculture ont pour but de faire vivre la population humaine mondiale en fournissant des produits alimentaires, des aliments pour animaux, de la bioénergie et des matériaux industriels. En effet l'enjeux de l'agriculture et de l'aquaculture à l'échelle mondiale, est de *fournir des ressources alimentaires, énergétiques et industrielles pour satisfaire la demande d'une population mondiale croissante* [180].

1.1.1 Les enjeux mondiaux des systèmes de production agricole et aquacole

L'objectif majeur de l'utilisation des méthodes en science des données dans les systèmes de production est l'amélioration de la qualité des produits [9]. L'amélioration de la qualité des produits est un aspect important dans la mise en place de stratégies commerciales pour les filières agricoles et aquacoles. Des paramètres techniques, tels que la taille, et/ou la croissance, ont un impact considérable, en fin de production, sur le prix des produits. A titre d'exemple, en aquaculture, la reconnaissance automatique des espèces permet d'extraire des descripteurs caractéristiques de la physionomie des espèces, et parfois de leur âge [78, 75]. En agriculture, pour les arbres fruitiers, l'extraction automatique de la taille des fruits, et de textures particulières d'une image, permet de préciser le moment propice pour la récolte [139, 4, 152]. L'état de maturation des fruits favorise une réduction des dépenses en termes de main d'oeuvre pour les opérations manuelles de récolte. La reconnaissance des espèces permet également d'identifier des caractéristiques après un avis expert, qui expriment la présence d'une maladie ou d'une déformation [126, 125, 47]. Les maladies et les parasites impactent également fortement les coûts de production et les bénéfices financiers des systèmes de production. L'identification automatique de leur présence assez tôt durant la période de culture permet de lutter contre leur expansion. La croissance des espèces est généralement associée à différent attributs temporels. Les paramètres comme la température, le pH (pour les sols), l'ensoleillement (pour les végétaux), sont des composantes essentielles pour analyser les mécanismes de croissance. De plus, pour lutter contre les maladies, le recours à des produits chimiques déversés directement sur les espèces et le sol, tels que des pesticides, impacte l'environnement. Il contamine le milieu de vie de ces espèces, mais aussi les eaux souterraines et d'autres écosystèmes locaux au travers de résidus. Le milieu d'élevage a un impact considérable sur la qualité de la production [140]. Dans les systèmes de production agricoles et aquacoles, l'analyse du milieu d'élevage reste essentielle pour être en mesure d'améliorer les rendements. Leur gestion nécessite à la fois de connaître leur influence sur les performances techniques, mais aussi leur évolution à différentes échelles de temps.

Il existe divers systèmes généralement associés à un niveau d'intensification et/ou de biodiversité destinés à la production de ressources naturelles. Néanmoins, la production intensive en monoculture est plus à même de modifier le milieu de culture et/ou d'élevage et de favoriser l'émergence de maladies.

La recherche, principalement dans le domaine agricole, et de manière moins répandue dans le domaine aquacole, a permis de mieux comprendre les facteurs à l'origine de ces maladies, d'anticiper leur apparition, et d'améliorer le rendement par la mise en place d'outils d'aide à la décision.

Pour des questions de santé publique, la chaîne d'approvisionnement des produits agricoles a fait l'objet d'une grande attention cette dernière décennie [46]. Les produits destinés à la consommation humaine (produits agro-alimentaires) sont soumis à des réglementations plus strictes et à une surveillance de plus en plus importante.

Des protocoles devant respecter des normes sanitaires de production des espèces, afin d'assurer leur qualité et leur traçabilité, imposent un suivi de toute la chaîne de production. De ce fait, d'importants moyens (en matériel informatique,..) sont mis en oeuvre pour obtenir des produits de qualité. Ces normes interviennent aussi en lien avec la gestion environnementale. La pollution produite par les exploitations, peut être mesurée, comme par exemple, dans le domaine de l'aquaculture, les rejets de matières organiques produits. Des réseaux de surveillance peuvent aussi être mis en place dans le domaine des pathologies. L'acquisition de données par images satellitaires est utilisée dans le domaine de la production afin d'avoir des données couvrant de grandes zones pour assurer un suivi à grande échelle [64, 110]. Le suivi de mesures pratiques permet aux systèmes de production d'avoir une meilleure compréhension du processus de production.

Ces différentes données produites conduisent à l'accumulation de différentes informations généralement complexes et bien souvent "cachées"..

1.1.2 Les données complexes dans les filières de production

Le données complexes sont dites complexes dans le sens où elles représentent une réalité complexe, sociale, environnementale, technique, décrite avec des points de vue multiples sur les objets composant le système, et à différentes échelles. La modélisation de cette réalité est définie dans un contexte spatio-temporel, dans lequel des objets sont comparables. Afin de déterminer un tel espace, les interactions ou les corrélations entre ces objets sont observées de manière empirique dans un premier temps, avant d'être modélisées. Le déroulement du processus de modélisation est fonction des causes et/ou des conséquences liées à ces interactions.

Les types de données à traiter pour répondre au enjeux en agriculture et en aquaculture :

Les données acquises dans l'agriculture et l'aquaculture, décrites dans la littérature, peuvent être de sources variées [110] telles que : des images satellitaires ou des images aériennes de drones, des données de capteurs (capteurs au sol : capteur de température, de salinité...), des données enregistrées manuellement, concernant par exemple la pratique d'élevage, ou encore des données textuelles relevées par l'exploitant, des données externes relevées par des stations météo etc...

Les données sont donc très diverses. De plus, à l'échelle d'une filière, certaines des données, en fonction de la qualité du matériel ou encore selon l'observateur, peuvent être imprécises, c'est-à-dire biaisées, ou ne pas être suivies régulièrement dans le temps. Ainsi les nouvelles méthodologies développées doivent tenir compte de l'imprécision de certaines données, mais aussi être en mesure d'analyser des séries incomplètes.

Dans les élevages agricole et aquacole, il y a des relations indirectes et surtout complexes entre 3 types de données :

1. **Les données concernant la production**. Ces données sont surtout statiques. Par exemple il y a le taux de mortalité ou encore la présence ou non de maladie en fin d'élevage. Elle peuvent être temporelles (évolution de la vitesse de croissance...).

2. **Les variables que l'on nommera variables forçantes** et qui se réfèrent à des données sur les espèces et l'environnement l'élevage sur lesquelles le fermier n'a pas d'influence. Ceux sont principalement des données statiques. On peut considérer par exemple que le climat, c'est à dire la température ambiante est une variable forçante, sachant que beaucoups de système de production se font en extérieur.

3. **Les données standards d'élevages.** Ces données standards peuvent être regroupées en deux catégories :

   (a) La variable de gestion : variable temporelle dont l'éleveur est en capacité d'agir directement sur la valeur à un instant défini.

   (b) La variable d'environnement : variable temporelle dont l'évolution est indépendante des pratiques d'élevage, et qui a une influence sur la qualité de la production.

Afin de représenter une réalité complexe, l'espace doit pouvoir intégrer différentes typologies des données. En effet, l'environnement dans lequel évolue les objets, leurs interactions, et les conséquences de ceux-ci peuvent être décrits par différents observateurs. L'impact de ces interactions est souvent étudié par des experts du domaine afin d'établir, à partir de données factuelles, des normes, sur une population représentative.

Ces normes doivent être représentatives des évolutions spatio-temporelles de ces interactions, afin de modéliser plus justement la réalité.

L'approche conventionnelle dans l'analyse pour la recherche de ces liens a recours à des méthodes statistiques sur des données produites en faible quantité et souvent de manière agrégée (moyenne des données relevées au cours de la production ou sur une période précise). Un test d'hypothèse est régulièrement utilisé afin de faire un choix entre deux hypothèses en comparant des résultats d'échantillons d'une population. Ces échantillons sont traduits en données numériques, liées, par exemple, aux attributs zootechniques, et aux paramètres physico-chimiques de leurs milieux de vie. Les données de ces tests font donc suite à une acquisition de données sur le terrain, sur des zones représentatives des différents aspects que peut avoir un système. Ces données sont aussi issues d'expériences visant à modéliser une réalité complexe. L'acquisition de données peut être très chronophage. Les testes d'hypothèses permettent donc de fournir une règle de décision, sur une base statistique comme l'écart entre la moyenne ou la médiane de ces échantillons. D'autres modèles visent à identifier les attributs les plus influents dans l'évolution, ou le lien entre ces différents paramètres. C'est une forme de réduction de la complexité des données. Enfin la statistique permet de définir une fonction permettant de modéliser l'évolution de ces paramètres réels par exemple dans le temps ou l'espace.

L'utilisation de ces méthodes conventionnelles n'est parfois pas suffisante pour décrire, et extraire des informations pertinentes d'une réalité complexe, comme celle que l'on observe dans une filière de production. Cette contrainte est due, en partie, à la typologie des données, mais aussi à la quantité exponentielle de données acquises.

En effet, les outils d'acquisition de données, dédiés au suivi des systèmes de production sont aujourd'hui de plus en plus variés. Il y a, comme nous le verrons, de l'acquisition par l'observation empirique, des relevées à l'aide de systèmes de mesures *in situ*, jusqu'aux observations à l'aide d'images satellites [147]. Il y a également des systèmes de surveillance vidéo intégrant des modèles capables de détecter des objets variés. Le format de données peut donc être très différent. Dans le cadre de la télédétection, l'utilisation d'images satellitaires est très courante pour le suivi des écosystèmes naturels sur de grandes surfaces [64, 110]. Les différentes fréquences d'acquisition des données par les capteurs entraînent la création de séries de données temporelles multi-échelles, pour lesquelles les méthodes statistiques conventionnelles ne proposent pas d'analyse multi-variée, multi-échelle adaptée (à leur typologie). Un autre exemple de typologie concerne les observations empiriques réalisées par les éleveurs au cours d'un élevage qui génèrent souvent des données plutôt textuelles et donc qualitatives.

Il est donc important de développer de nouvelles méthodologies, incluant des algorithmes d'apprentissage automatique [118], adaptés à la typologie de données complexes.

Dans les données complexes, les variables, qui décrivent les objets, peuvent aussi être imprécises spatialement, et temporellement. Cela signifie qu'elles peuvent être décrites de manières ambiguës. Elles sont parfois manquantes. Dans le cadre des systèmes de production, l'ambiguïté réside par exemple dans une description approximative de l'aspect physique de l'espèce en fin de production (moyen, assez grand...). Ces aspects qui seront détaillés par la suite, obligent les chercheurs en science de données à développer de nouvelles méthodologies et de nouveaux algorithmes pour la gestion des différents types d'imprécisions. Enfin, dans les données réelles, la complexité réside également dans les relations de dépendance entre les variables.

Dans certains jeux de données, le manque de données, de certaines variables, entraîne une forme de déséquilibre quantitatif entre elles. Ces déséquilibres impactent fortement l'apprentissage, puisqu'il se basera principalement sur les caractéristiques les plus présentes. Des approches de sélection de données peuvent les rééquilibrer en fonction de la quantité de ces caractéristiques par exemple. Les outils utilisés, qui sont décrits ensuite, et les protocoles choisis pour acquérir et enregistrer ces données, déterminent leurs qualités, en termes de potentiels descriptif et prédictif pour des algorithmes d'apprentissage automatique. Par exemple l'imprécision des capteurs ou encore le manque d'objectivité de l'observateur sont des facteurs qui génèrent de l'imprécision et impactent les performances d'apprentissage.

1.2 La science de données pour répondre aux enjeux du domaine agricole et aquacole

L'analyse de données complexes s'intéresse, par exemple, à la modélisation des relations et des variations spatio-temporelles entre des objets, appelés entités. Ces entités sont comparées selon des caractéristiques communes. Ces caractéristiques peuvent être analysées selon des mesures de distance ou de similarité adaptées à leurs typologies. Un attribut peut être décrit à différents instants dans le temps et dans l'espace, et selon différentes échelles spatio-temporelles. L'analyse de ces individus se fait, dans la plupart des cas, soit de manière exploratoire, soit en étant guidée, i.e supervisée par une ou plusieurs cibles à prédire. Elle peut également être effectuée par renforcement, c'est-à-dire de façon à optimiser une récompense quantitative au cours du temps. Cette thèse s'intéresse davantage à la création de nouvelles méthodes exploratoires, et utilisera les méthodes supervisées, à partir des nouveaux descripteurs

qui sont générés par les méthodes créées. Dans notre cas, l'objectif final est de créer des modèles d'apprentissage supervisés et adaptés aux données complexes du domaine d'étude. Elle permettra de créer des modèles prédictifs pour divers domaines. L'analyse exploratoire des individus se fait, dans le cas du clustering, par comparaison de ces individus. Comme énoncé, différentes métriques adaptées à la typologie des attributs seront utilisés pour les caractériser.

Il existe différentes approches d'apprentissage non supervisé qui utilisent des fonctions basées sur la densité ou la probabilité des valeurs des descripteurs afin de partitionner les individus. L'objectif est alors de les regrouper de manière homogène par rapport à ces valeurs, afin d'extraire de nouvelles connaissances utiles au domaine, grâce à une interprétation des groupes par des experts métier. L'homogénéité est définie selon différents critères. Dans le domaine agricole par exemple, ces critères peuvent être en lien avec des indicateurs (métier) sur la qualité du produit (couleur, dimension...). De manière plus répandue, l'apprentissage non supervisé vise à rechercher des classes en s'intéressant aux prototypes, c'est à dire à des individus représentatifs des groupes.

Il existe également un apprentissage non supervisé réduisant la dimension de l'espace de recherche. Les algorithmes d'apprentissage automatique permettent ainsi de créer des modèles descriptifs ou prédictifs. Ces modèles déterminent leurs résultats, et notamment leurs prédictions, en apprenant dans un premier temps à partir de données d'exemples, c'est-à-dire des données d'apprentissage. Pour créer de nouveaux modèles en science des données, adaptés aux données complexes générées par les systèmes de production, les interactions entre les performances de production, les conditions de production (environnementales, techniques) doivent être modélisées.

1.2.1 Le processus d'extraction de connaissance

Les méthodes se réfèrent à des algorithmes ou encore à un processus, incluant les algorithmes, permettant d'enregistrer, de préparer et d'analyser les données. Plusieurs algorithmes ont ainsi été développés pour assurer la mise en place du processus générique, appelé fouille de données, visant à extraire des connaissances utiles pour les expert métier. Le processus d'extraction de connaissances demande beaucoup de travail de préparation des données car les modèles utilisés doivent être robustes à la complexité des données, que nous définirons par la suite. L'extraction de nouvelles connaissances 'métier' permettra, par exemple, d'établir des normes de production. Ces normes aideront à définir des critères d'évaluation de la qualité des produits. Pour cela, une analyse exploratoire approfondie permettra de comprendre, dans un premier temps, la dynamique des données à différentes échelles spatio-temporelles avant de créer des modèles prédictifs de du système étudié.

La compréhension du domaine est primordiale, afin de déterminer une méthodologie d'analyse qui soit applicable aux données standards des filières de production. Des échanges avec des experts du métier doivent être organisés en amont du processus d'extraction de connaissances.

La figure 1.1 présente le processus itératif d'extraction de connaissances .

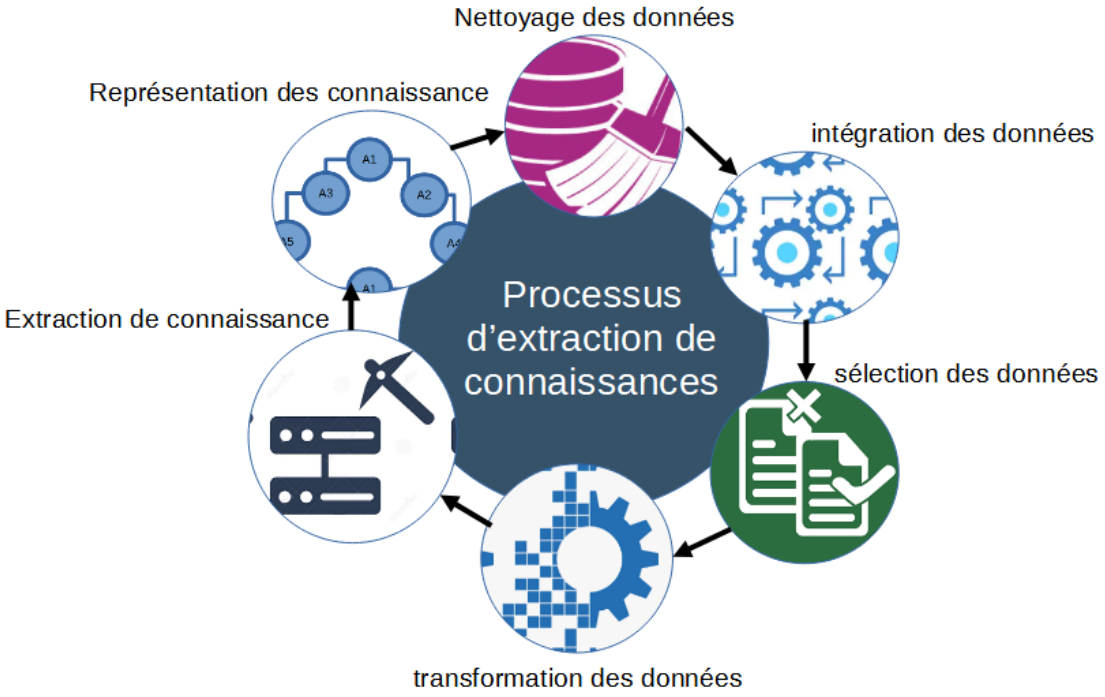

Fig. 1.1 Processus d'extraction de connaissances dans des données

Ce processus permettra d'intégrer les bases de données générées dans les systèmes de production, afin de créer des outils d'aide à la gestion des élevages. Notons quelques aspects indispensables de ces étapes :

- Le nettoyage des données concerne la suppression des données bruyantes et non pertinentes, et la rectification des données imprécises de la collecte.
- L'intégration des données dans une base de données normalisée assure un accès logique et rapide à ces données.
- La sélection des données se fait en lien avec l'expert métier, afin d'identifier les données pertinentes (avec un potentiel descriptif ou prédictif).
- La transformation des données permet d'obtenir des données au format adapté aux méthodes en science de données.
- L'extraction de connaissances utiles au métier se fait suite à l'interprétation, par l'expert métier, des résultats générés à partir de nouveaux modèles crées, ou en paramétrant des modèles existants.

- La représentation des connaissances doit faciliter une interprétation des résultats des modèles.

1.2.2 Les verrous scientifiques

L'un des principaux verrous scientifiques est de développer des algorithmes pour le croisement des données de différents types (statique, temporel, catégoriel) générés dans les systèmes de production et provenant de sources variées afin d'extraire de nouvelles connaissances utiles aux experts du domaines; Et d'ajuster le processus d'extraction de connaissances (figure 1.1) aux données complexes générées par les systèmes de production, en y intégrant de nouveaux algorithmes adaptés aux différents formats des données. On trouvera, pour ces systèmes, de nombreux travaux sur l'analyse de leurs données. La description de ces travaux sur un ensemble de filières sera fournie dans ce document. Ils concernent par exemple la modélisation mathématique et informatique de la dynamique de populations et des relations avec leur environnement. Nous nous intéresserons ici en particulier à la gestion de filières aquacoles. Notre approche devrait être valable et applicable pour différentes filières agricoles qui génèrent continuellement des données complexes. En raison des différentes sources possibles de ces données, cette thèse proposera une nouvelle méthodologie d'analyse de données pour les filières aquacoles, qui s'intègre logiquement au processus d'extraction de connaissances. L'objectif final est d'identifier des normes sur la variation des différentes variables et leurs liens au cours d'un élevage. L'enjeu est de comprendre, de décrire, de représenter, et de prédire les mécanismes complexes de cet écosystème réel, qui conduisent à une production de bonne ou de mauvaise qualité.

1.3 Plan de thèse

Nous nous intéresserons à l'utilisation des méthodes en science de données sur les données générées dans les systèmes de production aquacoles, principal domaine d'étude de cette thèse. Dans le chapitre 2, une description des méthodes les plus répandues est proposée, en y ajoutant un exemple d'application dans le domaine agricole. Le choix de ces méthodes mettra en évidence l'influence de la qualité des données d'apprentissage. Les méthodes seront présentées selon les deux grandes familles : l'apprentissage non supervisé et l'apprentissage supervisé par une ou plusieurs cibles associées aux individus;

La contribution de cette thèse est double : 1) elle proposera un nouveau processus de croisement de données générées dans les systèmes de production de ressources naturelles; 2) De nouveaux algorithmes pour le clustering de séries temporelles monovariées et multi-variées prenant en compte l'amplitude des séries, seront crées et intégrés à ce processus pour analyser l'évolution de la qualité du milieu d'élevage. Un nou-

veau formalisme définissant le clustering des séries temporelles multi-variées et multi-échelles sera proposé.

La chapitre 4 décrit en détail les nouvelles approches développées pour le clustering de séries temporelles mono-variées *Xmeans-TS* [167], et multi-variées *X-meansMMTS* [168]. Les méthodes seront testées et évaluées dans le chapitre de contribution, afin de prouver leur efficacité par rapport aux méthodes existantes. L'approche mono-variée sera comparée aux meilleures méthodes de clustering de séries temporelles monovariées sur une trentaine de jeux de données disponibles en ligne. L'approche mulitivariée est quant à elle plus complexe à comparer car aucunes méthodes existantes ne permet d'effectuer un clustering d'ensemble de séries temporelles multi-variées et multi-échelles. La performance de cette dernière méthode sera établie selon l'homogénéité des clusters à partir de plusieurs jeux de données de séries temporelles labélisées disponibles en ligne.

Les principales données analysées par les nouvelles méthodes proposées, sont celles de la filière aquacole et seront présentées dans le chapitre 4. Ces données sont générées dans la filière aquacole calédonienne. Ces données proviennent de différentes sources (ferme, usine de conditionnement...). Le chapitre 4 de présentation des données, mettra en avant leurs complexités (données statiques ou séries temporelles multi-variés, multi-échelles...).

La chapitre 5 présentera la stratégie d'analyse des données de la filière aquacole [166]. Elle comporte deux étapes dont la première étape sera décrite plus précisément dans le chapitre 6 et la seconde dans le chapitre 7.

Les données étudiées dans la première étape, sont liées aux principales variables influant sur les prix de vente et donc la rentabilité de la filière. De nouveaux paramètres de croissance, comme la vitesse de croissance initiale seront créés et analysées par des méthodes non supervisées. Ils serviront également d'attributs pour une analyse supervisée par des données de performance d'élevage. Ces paramètres de croissance serviront à identifier des périodes d'élevage lors de l'analyse de la qualité du milieu. Le chapitre 7, qui présente l'étape 2 de la stratégie d'analyse, fournira des résultats et une interprétation de l'analyse des séries temporelles liées à la qualité du milieu et croisée aux données de performance d'élevage. Une analyse mono-variée permettra de relever la corrélation entre les séries, prises indépendemment, et ces données de performance. L'interprétation des résultats de l'analyse mono-variée permettra de sélectionner des périodes d'élevage corrélées à la qualité des élevages et aux paramètres de croissance. Les données de qualité du milieu prisent à ces différentes périodes seront étudiées dans une analyse multi-variée.

Chapitre 2

L'intelligence artificielle pour la gestion des ressources biologiques

Ce chapitre est consacré principalement à la présentation des méthodes en science de données les plus appliquées, sur les données générées dans les domaines agricoles et aquacoles. Les variables standards d'élevages relevées dans l'introduction sont des données de qualité du milieu, des paramètres zootechniques, des données de performances d'élevage. Nous verrons comment ces données sont acquises, intégrées et analysées par les différentes méthodes développées dans le cadre du domaine des sciences des données.

Ces méthodes permettent en générale, de construire des modèles descriptifs et prédictifs. Plus particulièrement, elles font parties des méthodes d'apprentissage supervisé et non supervisé. Nous citerons ici, celles qui sont les plus répandus dans la littérature. Les avantages et les inconvénients de ces méthodes seront relevées pour des applications dans le domaine agricole et aquacole. Notons que l'aquaculture est devenue l'industrie majeure des produits de la mer [46]. Ce domaine est d'après la littérature, un domaine idéal qui a un besoin fort en outils d'aide à la décision pour bien gérer les ressources ..... Ces outils d'aide à la décision intègrent les modèles descriptifs pour comprendre les pratiques fermières en lien avec la production et la qualité des produits,et des modèles prédictifs pour gérer ces ressources et proposer des produits de qualité (fouille de données, apprentissage automatique, etc.) [79]. Le choix des méthodes dépend fortement de la nature des données, et en particulier de leur taille.Pour ce faire, nous allons dans un premier temps, donner une définition des données complexes et des typologies servant à modéliser la réalité complexe des systèmes de productions. Ensuite, nous présenterons les méthodes en science de données les plus utilisées en agriculture et en aquaculture.

2.1 Données complexes

Les données complexes sont caractérisées par des données hétérogènes, multi-variées, statiques, temporelles, multi-échelles. La description des phénomènes ou des objets complexes, peut se faire par un ensemble d'attributs. Ces attributs peuvent être de différents types.

Considérons un ensemble d'attributs hétérogènes $A = \{A_1, A_2, \ldots, A_p\}$, soit l'ensemble d'individus $I = \{I_1, I_2, \ldots, I_n\}$ possédant un espace de descriptions des attributs, noté X. Ces attributs sont aussi nommé 'descripteur'. Par exemple $I_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ où x_{ij} est la valeur de l'attribut j décrivant l'individu i. Les données sont dites complexes lorsque, par exemple, les descripteurs sont de types différents. Un individu peut être décrit par des descripteurs numériques ou non, i.e des données textuelles ou catégorielles. Par exemple l'attribut $A_1 \in \mathbb{R}$, $A_2 \in \mathbb{D}$ l'attribut $A_3 \in \{mauvais, bon, excellent\}$ peut être de type catégoriel et l'attribut $A_3 \in \{0, 1\}$ peut être de type binaire. Les données des individus peuvent évoluer dans le temps.

Dans le cas des données temporelles, la description des valeurs d'un attribut, est fournit à différents instants. Soit un ensemble d'instants $T = \{t_1, t_2, \ldots, t_p\}$, on considère une série de données aux différents instants $s = \{s(t_1), s(t_2), \ldots, s(t_p)\}$ où $s(t_1)$ est la valeur de la série s à l'instant t_1. Un individu $i \in I$ est représenté par sa série temporelle $s_i = \{s_i(t_1), s_i(t_2), \ldots, s_i(t_p)\}$.

A titre de comparaison, l'individu peut être représenté par un système produisant une quantité de ressources naturelles sur une période de temps bornée. Pour ces systèmes de production, cette période est dépendante des conditions climatiques qui auront une influence sur la productivité du système.

Des mesures de distances et de similarités, permettent de comparer des individus selon leurs attributs. Ces mesures sont adaptées à la typologie des attributs.

La distance entre individus : Les individus de l'espace X peuvent être comparés selon des mesures de distances ou de similarités, applicable aux *p* attributs qui les caractérisent.

Soit $Dist(I_i, I_j)$ la distance entre les individus i et j représentés par les séries s_i et s_j. En fonction de l'espace considéré, il est possible de définir la distance entre deux individus selon différentes approches. Considérons par exemple que l'espace X est un espace euclidien. Soit $Dist(I_i, I_j)$ la distance euclidienne entre l'individu i et j i.e $Dist(I_i, I_j) = \sqrt{\sum_{k=1}^{p}(s_i(t_k) - s_j(t_k))^2}$.

A partir de ces distances, un concept important, associé à la comparaison des données complexes et la notion de voisinage.

Le voisinage : Le voisinage d'un individu, détermine les individus qui sont proches, selon, par exemple, une mesure de distance adaptée aux types de données caractérisant les individus. Il est possible de déterminer une proximité spatiale par exemple à partir de la mesure euclidienne dans un espace euclidien. Soit X un espace euclidien et $Dist(I_i, I_j)$ la distance euclidienne entre les points I_i et I_j, le voisinage d'un point I_i (

$V_\epsilon(I_i)$), basé sur une distance ϵ peut être déterminé par :
$V_\epsilon(I_i) = \{Dist(I_i, I_j) < \epsilon | \forall j \in \{1, 2, \ldots, n\}, j \neq i\}$ l'ensemble des individus $I_j \in I$ tel que $j \neq i$, situés dans la boule ouverte centrée en I_i et de rayon ϵ. Dans le cadre des filières aquacoles, l'identification de structures de production avec des caractéristiques voisines passera par la comparaison des types de données générées dans ces filières que sont la qualité du milieu, la performance d'élevage, les données zootechniques, la qualité de production...

Nous verrons dans la section suivante différentes types de représentations de ces données complexes.

2.1.1 Représentation des données

Représentation matricielle des données Soit l'ensemble de n individus I = $\{I_1, I_2, \ldots, I_n\}$ et l'ensemble de p attributs A = $\{A_1, A_2, \ldots, A_p\}$

$$\text{Soit la matrice } X = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \ldots & A_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \ldots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \end{array} \text{de taille } p * n \text{ et Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

La ligne i de X est associée aux valeurs d'attributs de l'individus I_i. On notera $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ le vecteur de taille p lié aux valeurs observées qui décrivent l'individu I_i. Nous verrons dans certains cas que ces attributs peuvent expliquer, les valeurs d'une autre variable.

Cette variable à expliquer *Y* peut être de types variés tels que binaire ($y \in 0, 1$), entier ($y \in \mathbb{N}$), catégoriel i.e *y* est décrit selon des catégories.
La classification multi-label : Lorsque plusieurs cibles sont à expliquer, l'apprentissage supervisé est dit multi-label, comme ci-dessous :

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_1 & \ldots & A_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \ldots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \end{array} \text{et Y} = \begin{array}{c} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \ldots & Y_q \end{array} \\ \begin{pmatrix} y_{11} & y_{12} & \ldots & y_{1q} \\ y_{21} & y_{22} & \ldots & y_{2q} \\ \vdots & \ldots & \ldots & \vdots \\ y_{n1} & y_{n2} & \ldots & y_{nq} \end{pmatrix} \end{array}$$

Nous nous intéresserons plus loin aux méthodes développées et qui permettent de traiter avec des données multi-labels.

Représentation des données par un graphe : On appelle graphe G=(V,E) où V est un ensemble d'éléments appelés sommets ou nœuds et d'une partie *E* symétrique d'éléments e=(x,y), X*X appelés arêtes. Un nœud de G est étiqueté par un ensemble d'attributs A (numériques ou catégoriels). Chaque attribut $a \in A$ est associé à un domaine de valeurs.

Notations et définitions des séries temporelles multi-variées : Le formalisme courant des séries temporelles multi-variées, d'après la littérature, définit un ensemble de m séries temporelles $S = s_1, s_2, \ldots, s_m$, décrites sur un ensemble de temps avec : $s_i = s_i(t_1), s_i(t_2), \ldots, s_i(t_p)$. Ce formalisme est utilisé, par exemple, dans la recherche des sous séquences, dans les différentes séries, qui sont corrélées dans le temps. Une autre représentation permet de regrouper des individus selon un ensemble de séries temporelles multi-variées.

Par exemple, dans le cas de la représentation des données temporelles multi-variées à partir d'une matrice multidimensionnelle, les valeurs de la matrice peuvent être considérées comme des vecteurs. Pour chaque individu, chaque variable est une série temporelle. En reprenant l'exemple de la matrice M, l'individu $I_1 = (x_{i1}, x_{i2}, \ldots, x_{ip})$ devient une liste de vecteur ou $s_{ij} = (x_{ij}(t_1), x_{ij}(t_2), \ldots, x_{ij}(t_k))$ avec $x_{ij}(t_h)$ la valeur de l'attribut A_j de l'individu I_i à l'instant t_h. Et l'évolution temporelle des individus du graphe dans le temps est représentée par un ensemble de graphe (graphe dynamique), associé à l'ensemble de temps Dans le domaine agricole ou aquacole, l'individu peut être un élevage. Considérons un individu I_1, les données standards i.e comme la qualité du milieu (température, pH..), acquises au cours de l'élevage I_1, sont les attributs temporels A de la matrice. Notons que les différents attributs $A = \{A_1, A_2, \ldots, A_n\}$ peuvent être de tailles et d'échelles différentes. La complexité dans les systèmes de production est de définir le niveau de résolution le plus adéquat, et la meilleure période, qui assure une modélisation fiable entre ces séries de données et la qualité de production.

2.2 Les approches en apprentissage non supervisé les plus répandues en agriculture

Cette section présentera les algorithmes les plus répandus dans l'apprentissage non-supervisé : le clustering. Des exemples d'applications de ces méthodes dans le domaine agricole seront fournis.

Principe général de l'apprentissage non supervisée : L'objectif est d'extraire des groupes d'individus possédant des caractéristiques communes en les comparant selon

une distance adaptée à leurs attributs. Soit l'ensemble d'individus $I = \{I_1, I_2, \ldots, I_n\}$. On appelle k-clustering C de I, l'ensemble $C = \{C_1, C_2, \ldots, C_k\}$ contenant k sous-ensembles homogènes de I par rapport à une mesure de distance $Dist$ appliquées aux attributs. Chaque cluster $C_i \in C$ possède un représentant. Le représentant d'un cluster, appelé aussi prototype, peut être un centroïde, un médoïde, etc. noté R_{C_i} Un k-clustering $C = \{C_1, C_2, \ldots, C_k\}$ avec $\forall i \in \{1, \ldots, k\}$ et vérifie les critères suivants :

1. $X = \cup_{i=1}^{k} C_i$ et $C_i \cap C_j = \emptyset \; \forall i \neq j$.
2. $Dist(R_{C_i}, I) < Dist(R_{C_j}, I) \; \forall I \in C_i$ et $j \neq i$.

Les mesures de performance des méthodes non supervisées : Il est tout de même courant, d'après la littérature, de calculer la performance des méthodes non supervisées (clustering) à partir des classes connues d'appartenance des individus. Les classes peuvent être liées aux variables à expliquer. Ces classes peuvent être binaires ou catégorielles. Différentes mesures de performance des clusters , décrites ci dessous, se basent sur ces classes. Elles visent par exemple à déterminer l'homogénéité des clusters, sur la base des différentes classes associées aux individus qui les composent. Nous retiendrons les mesures qui suivent :

- **L'indice de Rand (RI)** [157] : est la probabilité pour que deux clusters soient en accord pour une paire de données choisie au hasard. Cette mesure est liée à la précision de la classification. R = $\frac{TP+TN}{(TP+TN+FP+FN)}$, où TP est le nombre de paires de séries qui appartiennent à la même classe et sont affectés au même cluster, TN est le nombre de paires de séries qui appartiennent à des classes différentes et sont attribués à différents clusters, FP est le nombre de paires de séries qui appartiennent à différentes classes mais qui sont attribuées au même cluster, et FN est le nombre de paires de séries qui appartiennent à la même classe mais sont affectées à des clusters différents.
- **L'indice de Rand ajusté (ARI)** [157] : l'ARI est Le score RI "ajusté" d'après le calcul suivant : ARI = (RI - $E(RI)$) / ($max(RI)$ - $E(RI)$) où E(R) est l'espérance de l'indice de Rand et max(RI) la valeur maximale qu'il peut atteindre.
- **L'homogénéité des clusters** [146] : Soit L un ensemble de labels réels $L = \{Li|i = 1, \ldots, m\}$, on désigne par a_{ij} le nombre d'instances de label i affectées au cluster j. L'homogénéité h calculée sur un ensemble de cluster C est définie

comme :

$$h(C) = \begin{cases} 1 & \text{If } H(L,C) = 0 \\ 1 - \frac{H(L|C)}{H(L)} & \text{else.} \end{cases}$$

où $H(L|C) = -\sum_{c=1}^{|C|}\sum_{l=1}^{|L|}\frac{a_{lc}}{N}\log\frac{a_{lc}}{\sum_{l=1}^{|L|}a_{lc}}$ and $H(L) = -\sum_{l=1}^{|L|}\frac{\sum_{c=1}^{|C|}a_{lc}}{m}\log\frac{\sum_{c=1}^{|C|}a_{lc}}{m}$

- **La complétude des clusters** [146] : La complétude mesure combien d'échantillons similaires sont regroupés par cluster : $c = \frac{H(L|C)}{H(L)}$

Notons que ces mesures, ne sont pas couramment utilisées pour déterminer la performance des méthodes d'apprentissage non-supervisées. Dans la littérature, l'homogénéité des clusters, est souvent calculée à partir d'une cible servant habituellement à superviser un apprentissage (pour des méthodes d'apprentissage supervisée). La cible, dans le cas, par exemple, des systèmes de production, peut être liée à la productivité du système. Les attributs peuvent être des indicateurs de qualité de l'environnement (qualité du sol ou de l'eau...). La cible est déterminée selon le métier. Et pour des données réelles, son choix est souvent discuté avec un expert du métier. Néanmoins il est toujours difficile d'obtenir une homogénéité supérieure à 50%, pour un apprentissage non supervisé (par la cible), même avec une cible corrélée aux attributs, puisque les approches ne sont pas guidées (par la cible) et que la pertinence de la cible doit être discutée au préalable avec des experts du domaine. L'homogénéité parfaite est donc obtenue lorsque chaque cluster ne contient qu'un individu. L'utilisation de ces mesures, dans cette thèse, visera surtout à comparer les nouvelles méthodes non-supervisées crées, avec les méthodes existantes. Nous verrons que malgré cette difficulté d'obtenir une homogénéité élevée, les nouvelles approches, assurent une amélioration de l'homogénéité, sur des données complexes (multi-échelles, multi-variées).

2.2.1 Le clustering de données statiques

Dans cette partie, les méthodes de clustering applicables aux données statiques seront présentées. Ces données statiques peuvent être catégorielles, numériques et nominales. Les approches adaptées aux séries temporelles seront ensuite exposées, sur la base des mesures de distance appropriées pour obtenir une distance minimale entre deux séries temporelles.

La méthode des k-means [65] : est une méthode de partitionnement qui découpe l'espace des données d'apprentissage X en k clusters disjoints. Soit $C = \{C_1, C_2, \ldots, C_k\}$ la partition de l'espace en k clusters. Considérons un espace euclidien. Pour obtenir ces clusters la méthode cherche à minimiser la distance entre un individu et le cluster auquel il doit appartenir :

$Kmeans(X) = Argmin(\sum_{j=1}^{k} \sum_{I_i \in C_j}^{n} (Dist(I_i, R_j)))$ k étant le nombre de clusters souhaité et :

- $Dist(I_i, R_j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - R_{jh})^2}$
- $R_j = \{R_{j1}, R_{j2}, \ldots, R_{jp}\}$ le barycentre de C_j avec $R_{ji} = \frac{1}{|C_j|} \sum_{h=1}^{p} x_{vh}$ où $x_v \in C_i$ et h $\in \{1, 2, \ldots, p\}$

Algorithm 1 pseudo code algorithme de Kmeans
Input:

1: Placez aléatoirement k centroïdes sur les données
2: **while** pas de convergence **do**
3: attribuez les observations au centroïde le plus proche.
4: Recalculez les centroïdes
5: **end while**

L'algorithme est utilisé pour divers objectifs dans l'agriculture et notamment dans l'analyse d'images, pour leur segmentation. A titre d'exemple, cette approche s'inscrit dans l'objectif de répertorier les différents types de diversité des plantes et de réaliser pour cela des bases de données afin d'exploiter leurs propriétés thérapeutiques. Les plantes ont besoin de plusieurs nutriments (phosphore..) pour se développer. Le clustering sur les valeurs de pixels permet d'extraire des objets en vue d'étudier la variance des valeurs. Cela assure la reconnaissance de variations particulières de couleurs qui conduisent à la reconnaissance de carence en nutriments à partir de l'aspect des feuilles [153].

La définition de la valeur optimale des clusters est essentielle dans l'algorithme *k-means*. Or son choix est souvent empirique.

Dans [161], la valeur optimale des clusters est déterminée par différentes approches qui incluent la méthode du coude [129]. Les auteurs analysent des performances des algorithmes *K-means* et *K-medoid* sur des données agricoles. La différence entre *K-means* et *K-medoid* [87], réside principalement dans le choix du représentant, qui est le point le plus centrale pour la méthode *K-medoid*. Cette différence modifie considérablement les performances en fonction des données, en faveur de *K-medoid*. Dans cette thèse, la nouvelle méthode proposée, pour le clustering de séries temporelles génère des représentants caractérisant mieux les individus que les méthodes existantes. Cette amélioration est due à une nouvelle mesure de dispersion qui applicable à la méthode *K-means*.

Le clustering flou [158] L'approche dite floue par la méthode *Fuzzy C-means* est très similaire à la méthode K-means. La distance entre les instances et les centroids, se base, à la différence de *K-means*, sur un degré d'appartenance aux clusters. Ce degré d'appartenance est déterminé en fonction des intervalles de valeurs qui peuvent être les distances euclidiennes, par exemple, entre les individus et le centroide du cluster. L'algorithme s'arrête lorsque que la variation des degrés d'appartenance entre deux itérations ne dépasse pas un seuil déterminé.
Il existe plusieurs approches dites floues, utilisant différentes mesures de distance. Avec cette méthode, [121] analysent à partir de méthodes de clustering floues, les données collectées dans les étables à litière de composte. Ces étables sont des systèmes de confinement moderne, où les bovins ont plus de liberté de mouvement à l'intérieur de l'installation et peuvent se coucher de manière plus 'naturelle'. Les classifieurs flous ont été développés pour aider à la prise de décision pour le contrôle des variables telles que l'humidité, la température et l'aération de la litière. L'idée est de promouvoir le bien-être du bétail, et d'améliorer les indices de productivité. Des données provenant de 42 étables dans l'État du Kentucky aux États-Unis, ont été prises en compte. Ainsi dans [121], six classes liées aux degrés d'efficacité du processus de compostage ont été identifiées.

La méthode DBSCAN décrite par [45] est basée sur la densité des points. Il se sert pour cela de deux paramètres pour identifier les points à intégrer dans un cluster. Le premier est la distance ϵ dans laquelle doit se trouver le nombre minimum de points $minpoints$ qui correspond au deuxième paramètre. Pour chaque point, si son ϵ-voisinage contient le nombre minimum de points paramétrés, alors ils font partie du même cluster. Par itération la méthode *DBSCAN* parcourt le ϵ -voisinage de proche en proche pour obtenir l'ensemble des points du cluster.

L'approche est régulièrement utilisé pour l'analyse spatiale afin d'agglomérer les régions, ou les cultures selon le type de sol. Dans [109], l'exploration de données par *DBscan* est appliquée aux besoins de fertilisants, en classant la fertilité des sols. Les classes sont obtenues sur la base des valeurs de nutriments présents dans les sols. Un modèle de fertilisation par groupe est appliqué, en considérant les valeurs agrégées des nutriments dans les groupes.

Les paramètres d'entrées sont donc une estimation de la densité de points des clusters.
En vue d'extraire des informations pertinentes pour l'agriculture, des observations météorologiques ont été analysés par [37] par la méthode *DBSCAN* afin de les prédire. Pour cela des données atmosphériques sont enregistrées toutes les heures, et sont transformées en base de données structurelles sur laquelle la méthode *DBSCAN* est appliquée. Ensuite, le protocole basé sur la priorité est utilisé sur les clusters résultants pour donner la

prédiction météorologique à partir des données collectées au cours des trois dernières années. Cette prédiction a un intérêt pour l'agriculture, qui dépend fortement de variables environnementales (forçantes) comme le climat. En effet la température impacte certains paramètres zootechniques des espèces élevées et/ou cultivées.

Le clustering hiérarchique [81] définit à l'état initial chaque individu comme son propre cluster. De manière itérative, l'approche ascendant combine les paires de clusters les plus similaires, jusqu'à ce que tous les points soient dans le même cluster. On obtient un dendrogramme avec des niveaux de combinaison de paire de clusters, intégrant au fur et à mesure des successions de combinaisons, de plus en plus d'individus, issus des regroupements précédents. L'approche descendant vise à découper l'espace par exemple par dichotomie selon une mesure de distance adaptée aux valeurs.

L'une des tâches clés dans le domaine de l'agriculture est la délimitation des zones de production et de gestion. Dans [147], les auteurs utilisent une variante du clustering hiérarchique en conjonction avec une contrainte spatiale. En considérant un ensemble de données géo-référencées provenant d'image à hautes résolutions spatiales, l'objectif de ces auteurs a été d'extraire des zones (pixels) contiguës possédant des caractéristiques similaires, et de différencier des zones possédant des caractéristiques différentes.

Son approche a donc été d'utiliser une variante du regroupement spatial (hiérarchique) sous la contrainte de conserver les groupes résultants spatialement contigus.

L'objectif de délimiter géographiquement des élevages dans le domaine aquacole reste important, mais très peu étudié.

Algorithm 2 : pseudo code algorithme de clustering hierarchique
Input:

1: Définir chaque individu comme étant une classe.
2: Calculer la matrice des distances des individus 2 à 2
3: **while** tous les individus ne sont pas regroupés en une seule classe **do**
4: Regrouper les 2 éléments (individus ou groupes) les plus proches au sens d'un critère chosie.
5: Mettre à jour la matrice des distances en remplaçant les deux éléments regroupés par le nouveau et en recalculant sa distance avec les autres classes.
6: **end while**

2.2.1.0.1 La carte auto-adaptative (Self-organizing map) La carte auto-adaptative [93] est une méthode de clustering basée sur les réseaux de neurones artificiels. Dans la sous section suivante, il y a une description de ces réseaux habituellement utilisés pour un apprentissage supervisé. les neurones de la carte auto-adaptative, encore appelée carte de *Kohonen*, sont positionnés de manière arbitraire dans l'espace des données.

L'apprentissage permet d'obtenir une cartographie des neurones qui tend à se rapprocher de la répartition de données dans un espace à grande dimension. Le nœud le plus proche de la donnée d'apprentissage est sélectionné. Il est déplacé vers la donnée d'apprentissage, comme le sont (dans une moindre mesure) ses voisins sur la grille. Après plusieurs itérations, la grille tend à se rapprocher de la distribution des données réelles.

2.3 Les méthodes de classification supervisées les plus répandues

2.3.1 La classification supervisée monolabel

Principe général de l'apprentissage supervisé : Dans la classification supervisée, les objets sont associés à une ou plusieurs étiquettes qui seront les données ciblées lors de l'apprentissage.

Cette classification vise à modéliser la relation entre les individus, décrits dans l'espace X, et les étiquettes Y par la fonction $f(X) = Y$. Comme énoncé précédemment, les données acquises d'un système, sont dans la réalité régulièrement bruitées. Elles contiennent des erreurs dues à l'acquisition (l'imprécision des capteurs, le manque d'objectivité de l'observateur...). La classification supervisée cherchera par approximation à déterminer la fonction avec une erreur ϵ i.e $f(X) = Y + \epsilon$. Le but est d'ajuster la précision pendant le processus d'apprentissage, en limitant l'écart entre le résultat attendu et le résultat fournit par l'algorithme i.e l'apprenant. Rappelons que, par rapport à l'apprentissage, des données temporelles interviennent dans la modélisation d'un système de production. Ainsi la complexité de l'apprentissage, et notamment de l'apprentissage à partir de variables temporelles qui varient fortement au cours du temps, est d'ajuster la précision en fonction de la variation des séries.
Dans le cas d'une classification de systèmes de production, les cibles correspondent aux données de productivité et/ou de qualité des produits, qui concernent par exemple des paramètres physiologiques sur des espèces élevées.

Les mesures de performances des méthodes supervisées: La prédiction du label d'un ou de plusieurs individus, est l'objectif principal de la classification. Pour cela, la réponse émise par l'algorithme d'apprentissage, est comparée aux données réelles, selon son type de données. L'étude peut être souvent portée sur un phénomène de nature binaire qui sera donc le type de données de la cible. Par exemple, dans le domaine agricole, l'algorithme peut déterminer sur la base d'une image la présence ou l'absence d'une maladie. Dans le cas d'une classification binaire, les quantités suivantes (présentées dans [119]) sont utilisées pour évaluer la performance :

- Vrais positifs (VP) : nombre d'individus réellement positifs et déclarés également

positifs par l'algorithme,

- Faux positifs (FP) : nombre d'individus déclarés positifs mais qui sont en réalité négatifs,
- Vrais négatifs (VN) : nombre d'individus réellement négatifs et déclarés également négatifs par l'algorithme,
- Faux négatifs (FN) : nombre d'individus détectés négatifs mais qui sont en réalité positifs

À partir de ces éléments les mesures suivantes (présentées dans [119]) sont déterminées :

- L'*accuracy* indique le pourcentage de bonnes prédictions $\frac{VP+VN}{VP+VN+FP+FN}$
- La *précision* est le pourcentage de prédictions correctes parmi les prédictions positives $\frac{VP}{VP+FP}$
- la *sensiblité* ou le rappel est le pourcentage de positifs réels prédit correctement $\frac{VN}{VN+FP}$ (taux de VP)
- la *spécificité* est le pourcentage de négatifs réels prédit correctement (1 - sensibilité)

La reconnaissance automatique d'espèces ou de maladies, à partir d'images, en fonction des caractéristiques de l'image et notamment des bandes spectrales, font souvent l'objet de cibles binaires à prédire [126, 125, 47] . Par exemple sur la figure 2.1, l'algorithme reconnaît avec une précision de 100% l'appartenance à la classe 'C5', qui admet la présence d'une maladie en fonction de la disposition d'une couleur jaunâtre sur les feuilles de bananier.

Les types d'apprentissages en profondeur (deep learning), décrits plus bas, sont des réseaux de noeuds connectés qui au final émettent une réponse en fonction des caractéristiques de l'image ; On retrouve, à partir des données présentées ci-dessus, des performances d'*accuracy* élevées. Ces scores peuvent être supérieurs à 95% de précision pour la reconnaissance des maladies liées aux plantes (cf figure 2.1).

Pour des cibles non binaires, d'autres mesures peuvent être utilisées; Soit n le nombre d'individus, et $\hat{y}$ la prédiction de l'algorithme. La performance des modèles apprenant sur des cibles non binaire est calculée à partir des mesures suivantes :

- l'erreur quadratique moyenne (*mean squared error MSE*) : $\frac{1}{n}\sum(y-\hat{y})^2$
- l'erreur quadratique absolue (*mean absolute error MAE*) : $\frac{1}{n}\sum|(y-\hat{y})|$

Fig. 2.1 Exemples représentatifs d'une classification pour la présence d'une maladie avec une performance supérieure à 95%. (d'après [47])

Les modèles bayésiens : Le modèle bayésien [144] classifie des données labélisées en fonction des labels ayant la probabilité conditionnelle la plus élevée. Suite à cela, un réseau bayésien est crée, et est capable de représenter l'impact ou l'influence des informations sur les données existantes par des expressions probabilistes décrivant les relations entre les variables. Les relations peuvent être représentées graphiquement par des structures de graphe. La règle de Bayes permet de définir les probabilités postérieures liées à chaque classe. Ces probabilités sont calculées à partir de la probabilité conditionnelle, $P(x_j|y_i)$ et de la probabilité à priori $P(y_i)$ comme suit : $P(y_i|x) = \frac{p(x|y_i)*P(y_i)}{P(x)}$ où $p(x) = \sum_{i=1}^{n} p(x|y_i) * P(y_i)$
Ce théorème peut être utilisé pour mettre à jour ou réviser les degrés de croyances des probabilités des états variables, sachant certaines valeurs, et à la lumière de nouvelles informations.
[40] proposent une revue complète de l'application des réseaux bayésiens dans l'agriculture, affirmant que ces réseaux sont particulièrement adaptés à la recherche agricole en raison de leur capacité à raisonner avec des informations incomplètes et à intégrer de nouvelles informations.

Par exemple, [16] se servent de l'approche de *Bayes* sur des données empiriques recueillies dans une culture de maïs, afin de voir l'impact des espèces concurrentes. Les caractéristiques utilisées pour construire le classificateur de réseau bayésien sont la densité totale des mauvaises herbes et les proportions correspondantes de narcisses et de dicotylédones. La classe est la biomasse des mauvaises herbes.

L'arbre de décision [133] : est une méthode d'apprentissage supervisée, qui permet de déterminer un ordre de test à effectuer sur les attributs A décrivant les objets I afin d'obtenir une classification automatique dans l'espaces des cibles Y. Cette méthode divise l'espace X en sous espaces selon un critère qui définit l'ordre de sélection des attributs pour la subdivision de l'espace. Pour cela, des tests sont appliqués aux at-

tributs. Les tests visent à déterminer l'attribut qui permet d'optimiser le plus le critère. L'attribut qui optimise le plus le critère est sélectionné et retiré ensuite de la liste des attributs. Le processus est itéré jusqu'à que l'ensemble des attributs soit sélectionné. L'algorithme 3 présente un pseudo code de la construction de l'arbre de décision.

Algorithm 3 Arbre de Decision

Input:

- exemples = liste d'exemples étiquetés
- questions = liste des attributs non utilisés jusqu'à présent

Output: - $Graphe$ arbre de décision

```
1:  if question vide then
2:      retourner feuille(T)
3:  else
4:      q = attribut optimal (avec le plus grand gain d'entropie)
5:      n = nouveau nœud créé qui testera l'attribut q
6:      for chaque v = valeur possible de q do
7:          e = l'ensemble des éléments des exemples ayant v comme valeur à l'attribut
    q
8:          chaque nœud fils de n est créé par ID3(e, questions - {q})
9:      end for
10:     Retourner n
11: end if
```

L'une des méthodes les plus répandues est l'algorithme *C4.5* qui détermine la structuration et l'ordonnancement des noeuds (i.e attributs) en fonction du gain d'information. Il existe différents critères permettant d'ordonnancer les attributs.

La classification se sert de cette structure pour déterminer le chemin, de la racine (l'attribut le plus optimal) vers la feuille (la cible prédite), en fonction des valeurs d'attributs des individus à classer.

L'approche peut être utilisée dans l'agriculture pour ordonnancer l'influence des variables environnementales et de gestion sur la productivité des systèmes et la qualité des produits.

Dans [113], l'arbre est utilisé pour plusieurs facteurs qui affecteraient le Produit Intérieur, Brut (*PIB*) agricole. L'auteur se sert comme attributs des données statiques, notamment en lien avec les données environnementales, et d'autres variables comme par exemple, la population agricole, le nombre de têtes de bétail, la température moyenne, la durée d'ensoleillement ... Ces attributs ont permis de déterminer les principaux facteurs et la manière dont ils affectent la valeur (en terme de *PIB*) de la production agricole.

La figure 2.2 présente l'arbre et les variables utilisées, ordonnées selon leurs impactes dans le *PIB*. Les feuilles de l'arbre sont donc les cibles catégorielles (i.e l'impacte

fort ou faible).

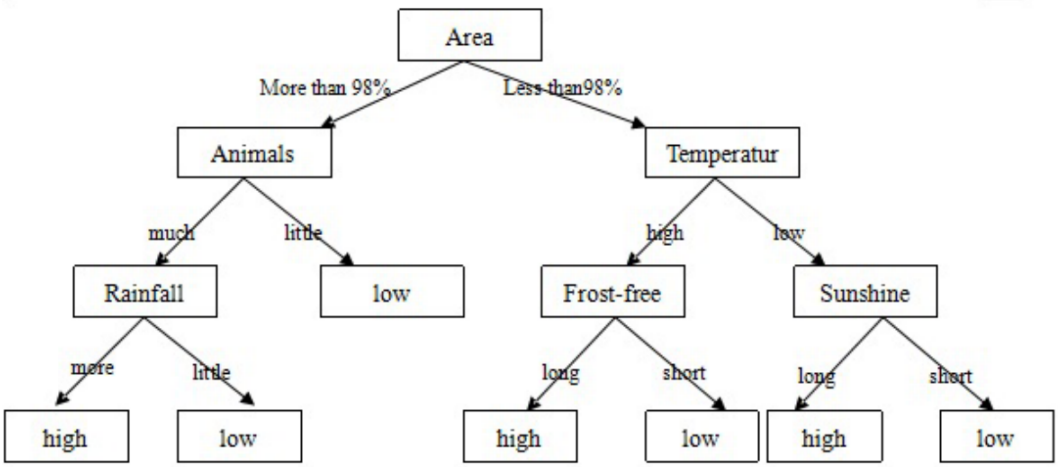

Fig. 2.2 Structure de l'arbre de décision pour déterminer la valeur de la production agricole (Produit Intérieur Brut Agricole) ([113])

Les règles obtenues pour la classification sont par exemple :
1) **SI** la superficie est supérieure à 98% et que les animaux ont des litières, **ALORS** le rendement est faible.
2) **SI** la superficie est supérieure à 98% et qu'il y a beaucoup d'animaux et que les précipitations sont faibles **ALORS** la production est faible

Autre exemple dans [179], les auteurs traitent principalement d'un classificateur spécifique à base d'un arbre de décision, utilisé pour classer des données agricoles. Il est capable de traiter à la fois des données complètes comme incomplètes. La méthode est appliquée à plusieurs ensembles de données. Le premier est utilisé pour prédire l'âge de l'ormeau à partir de mesures physiques. Ces mesures concernent des données zootechniques comme la longueur, le diamètre et le poids de la coquille.

Les séparateurs à marge (SVM) Les séparateurs à marge [172] sont des classifieurs supervisés qui déterminent une frontière de décision pour discriminer les objets dans l'espace d'apprentissage X. Cette frontière est déterminée par la recherche d'un hyperplan $h(x) = (w^T x + w_0)$ séparant des objets selon les labels Y avec $w^T x$ le vecteur normal à l'hyperplan. Cet hyperplan optimal est déterminé par : $Argmax_{w,w0}$ min$\{||x - x_i|| : x \in \mathbb{R}^d, h(x) = 0, i = 1, \ldots, n\}$, $h(x)$ est déterminé en maximisant les distances des points les plus proches de sa norme w en les considérant par classe (par label).
Il existe différentes approches permettant de déterminer la frontière. La résolution de l'équation $h(x) = (w^T x + w_0)$ est obtenue selon différentes approches par exemple en

changeant la forme de l'équation.
Dans le domaine agricole, cette approche est très répandue. C'est une stratégie d'apprentissage automatique qui est concurrente aux types d'apprentissages précédents et notamment aux modèles bayésiens. [94] présentent l'application du *SVM* dans l'agriculture, pour aller vers une meilleure gestion des cultures. Le *SVM* donne de bonnes approximations pour des échantillons et des dimensions faibles. Il reste cependant très sensible à la qualité des données. Les objectifs en analyse d'images sont très variés et comprennent la détection de maladies à partir des caractéristiques visibles sur l'espèce cultivées (feuilles de riz) . La prédiction du rendement est aussi un enjeu pour lequel l'approche *SVM* est utilisée. Par exemple, [64] expérimente l'approche sur des données d'images pour prédire le rendement du blé, à partir des données issues des bandes spectrales d'images satellitaires. Ils utilisent notamment l'indice normalisé de végétation de différences (*NDVI*). Pour cela la prédiction a été faite sur différentes fenêtres temporelles liées à différentes périodes de croissance de la plante.

Les k plus proches voisins : La méthode des *K* plus proches voisins (cf algo 4) [28] est une méthode non paramétrique qui détermine la classe d'une observation en utilisant une valeur moyenne, où la classe la plus commune, de ces points *k*-plus proches. Le choix de la distance à utiliser et du nombre de voisins *K* à considérer peut ne pas être évident. Afin d'obtenir une performance correcte, par cet algorithme, il est souvent nécessaire souvent de tester l'approche avec différentes combinaisons.

Algorithm 4 k plus proches voisins

Input:
- $X = \{x_1, x_2, \ldots, x_p\}$
- Y une cible quelconque
- K le nombre de voisins à considérer

1: **for** chaque point des données de test **do**
2: Calculer la distance (euclidienne) de tous les points des données d'apprentissage
3: Stocker les distances dans une liste et la trier
4: Choisissez les k premiers points
5: Attribuer une classe au point de test en fonction de la majorité des classes présentes dans les points choisis.
6: **end for**

L'approche est adaptée aux problèmes non linéaires, et peut être sur des données complexes, plus performantes que l'approche *SVM*. La méthode des *k plus proches voisins* est moins répandue dans l'agriculture, mais à néanmoins fait ses preuves dans l'analyse des coûts de fonctionnement des systèmes de production. Dans l'étude de [11], l'objectif a été de développer un outil de planification permettant aux agriculteurs d'évaluer les systèmes logistiques, la finalité étant de réduire les dépenses liées

à la consommation d'énergie, en optimisant l'apport de ressources, notamment pour différents types de véhicules tracteurs. Cela permet aux agriculteurs de connaître les paramètres clés des différents systèmes, tels que la consommation moyenne de carburant et la vitesse moyenne de transport. Les données concernaient des trajets en Allemagne, comprenant des routes transversales et des routes de campagne.

De manière analogue, le principal objectif de [3] était de déterminer la capacité de l'algorithme des *k voisins les plus proches* (KNN) à prédire correctement la consommation de carburant des systèmes "tracteur-chisel-charrue". La méthode (sur des données non-linéaires) présentait des coefficients de corrélation sur l'ensemble des données de 0,817. Cette même performance était de 0,422 en utilisant la méthode de régression linéaire multiple.

L'algorithme a été utilisé par [162] dans le cadre d'une analyse spatiale. Afin d'aider à déterminer le type de cultures alimentaires qui convient à une parcelle, l'auteur développe un système de méthodes d'identification des terres en utilisant la méthode *KNN*. L'étude a été conduite en Indonésie, où la plupart des habitants sont des agriculteurs et travaillent dans le secteur agricole. Les attributs utilisés concernaient les paramètres édaphiques, et météorologiques (température, précipitations (mm/année), taux d'humidité, drainage, texture...). Ces paramètres impactent l'espèce cultivée. Dans l'aquaculture le positionnement des bassins est aussi un facteur prépondérant dans le développement de système de production d'espèces aquatiques; Les paramètres édaphiques impactent, le milieu d'élevage, de manière analogue à l'agriculture.

Les réseaux de neurones artificiels: Un réseau connexionniste [68] est une structure dans laquelle des neurones sont connectés en couches. Ces réseaux propagent des valeurs d'activations entre les couches selon une règle qui permet de déterminer la valeur en sortie par neurone. Nous ne nous attarderons pas sur ce concept, il reste néanmoins l'un des plus performants dans les approches supervisées.

Par exemple, le schéma de la figure 2.3 représente un perceptron. A droite un signal de sortie est véhiculé en fonction d'un seuil d'activation. Ce seuil est comparé à la valeur obtenue par le noyau 'combinaison' qui effectue un traitement sur le vecteur d'entrée x. Chaque composante x_i est pondérée par le poids synaptique correspondant w_i, en appliquant une fonction tel que : $\sum_{i=0,d} w_i * x_i$. avec dans cette exemple $d = 4$

L'un des réseaux les plus utilisés en agriculture est le réseau de neurones convolutifs (*CNN*, [86]), qui permet d'extraire des caractéristiques particulières liées à un ou plusieurs objets dans l'image. Dans ce réseau, des filtres sont appliqués à l'image successivement, pour en créer des couches distinctes. Un filtre peut par exemple permettre d'extraire les valeurs maximales, de plusieurs groupes de pixels pris à différents emplacements dans l'image pour les passer à la couche suivante.

Les techniques dites floues permettent de prendre en compte l'imprécision des données

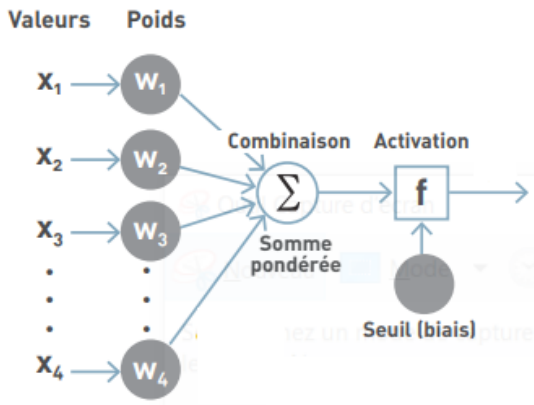

Fig. 2.3 Schéma d'un perceptron.

de plusieurs manières, à partir d'ensembles flous qui sont des ensembles dont les éléments sont décrits par des classes. Pour rappel : les données imprécises peuvent faire référence à des données d'apprentissage dont la description des objets est vague. Une description théorique des modèles principalement connexionistes combinés à ces techniques est présentée ci-dessous. Nous présenterons ici succinctement ces méthodes basées sur des réseaux de neurones artificiels :

- **Radial basis function network (*RBFNN*)** [17] : Réseau particulier à base radiale ou la réponse diminue ou augmente par rapport à un point central. Les fonctions radiales de base peuvent êtres des fonctions gaussiennes, multi quadratiques....
- **Fuzzy neuronal** [95]: On peut considérer dans ce cas les entrées réelles et des poids flous, les entrées floues et de poids réels, sinon les deux peuvent être flous. Ce système n'est pas très répandu en raison du manque de données qui impliquent ces 3 modalités.
- **le réseau de neurones récurrents** [67] : Un réseau pour lequel il existe au moins un cycle. Il est adapté aux données séquentielles c'est à dire de signaux. Les connexions récurrentes permettent d'analyser la partie passée du signal, i.e de mémoriser sa décision.

Les réseaux combinés à la logique floue intéressent particulièrement la télédétection. Les pixels de basses résolutions ($>$ 1m de résolution), contiennent des informations hétérogènes. Dans le cadre d'une étude avec des images sur des parcelles agricoles, les pixels peuvent contenir une variété de plantes en fonction de la résolution du pixel. La complexité de l'information de pixels mixtes impacte la performance des algorithmes d'apprentissage. En effet, dans la grande majorité des articles d'apprentissage avec des données d'images, la classification est basée sur le pixel. Dans le cadre d'une classification mono-label, chaque pixel est associé à une classe, afin d'assurer l'apprentissage des

réponses spectrales de l'image par rapport à la classe, qui peut être le type de plante. Or chaque pixel peut être composé d'adhésions multiples et partielles de toutes les classes candidates. La logique floue permet d'attribuer pour chaque pixel, un degré d'appartenance à différentes classes. [123] font une review de la logique floue appliquée, à la classification des cultures par imagerie.

Utiliser les approches floues permet de déterminer des 'labels" cohérents avec l'expérience métier, et de représenter des cibles complexes.

2.3.2 Apprentissage multilabel

Les approches d'apprentissage multi-label peuvent être définies selon trois grandes familles [169, 184, 6] :

1. Approches d'apprentissage par transformation : elles transforment le problème d'apprentissage multi-label en un ou plusieurs problèmes de classification ou de régression mono-label, relationnelles ou non.

2. Approches d'apprentissage par adaptation : elles adaptent des algorithmes d'apprentissage pour des données multi-label,

3. Approches d'apprentissage ensemble : elles utilisent un ensemble de classifieurs issus de la première ou de la deuxième famille d'approches.

Les approches par transformation : Binary Relevance [114] *BR*, est la méthode la plus populaire et la plus simple de cette classe d'approche. Elle transforme le problème d'apprentissage multi-label en q problèmes de classification ou de régression mono-label. Pour l'apprentissage de chaque label $Y_i (1 \leq i \leq q)$, un classifieur binaire est utilisé. *BR* retourne l'union des prédictions de chaque classifieur. Classifier Chain *CC* [143] est une amélioration de la méthode *BR* (qui transforme également le problème d'apprentissage multi-label en q problèmes de classification ou de régression mono-label).. Chaque classifieur binaire apprenant un label Y_i ajoute, dans son espace d'attributs, tous les labels associés aux classifieurs qui le précèdent dans la chaîne (i.e. $Y_1, \ldots,$ Y_{i1}). Cependant, les classifieurs sont entraînés dans un ordre aléatoire défini avant la phase d'apprentissage $[1.., i, ..q]$.

Les approches par adaptation : Dans cette approche les algorithmes sont modifiés. Prenons le cas Multi-Label, kNN (ML-kNN) [183] est une méthode de type Binary Relevance *BR* qui combine l'algorithme standard de kNN avec une inférence bayésienne. En phase d'apprentissage, *ML-kNN* estime les probabilités a priori et a posteriori de chaque label à partir des exemples d'apprentissage. Pour un nouvel exemple I_i , *ML-kNN* calcule ses *k* plus proches voisins, il mesure la fréquence de chaque

label dans ce voisinage. Cette fréquence est ensuite combinée avec les probabilités estimées dans la phase d'apprentissage pour déterminer son ensemble de labels en suivant le principe du maximum a posteriori (MAP). *ML-kNN* a l'avantage de tirer parti du raisonnement bayésien : la frontière de décision peut être ajustée de manière adaptative dûe à la variabilité des voisins des nouveaux exemples et le problème du déséquilibre entre classes peut être largement atténué à l'aide des probabilités a priori estimées pour chaque label. Néanmoins, comme toute approche de type *kNN*, le temps de prédiction croît linéairement avec la taille de l'ensemble d'apprentissage.

Les approches par apprentissage d'ensemble : Ensemble de Binary Relevance (*EBR*) et Ensemble de Classifier Chains *ECC* [143]) entraînent respectivement *N* classifieurs *BR* et *N* classifieurs *CC*. Pour diversifier les classifieurs, elles sous-échantillonnent de façon répétitive (avec remise) l'ensemble d'apprentissage (i.e. bagging [15]). Chaque classifieur *CC* est entraîné suivant un ordre de label différent défini aléatoirement. *EBR* et *ECC*. L'avantage de ces deux méthodes d'ensemble est qu'elles tentent d'améliorer les performances de leurs classifieurs de base (*BR* et *CC*) en multipliant les modèles. Néanmoins, leur complexité d'apprentissage croît linéairement avec le nombre de labels.

2.4 L'utilisation des méthodes en intelligence artificielle pour répondre aux enjeux et aux problématiques dans le domaine aquacole

Il y a des objectifs communs, traités par les méthodes en science de donnée, entre le domaine agricole et le domaine aquacole. Comparativement aux productions agricoles, la littérature contient beaucoup moins d'articles dans le domaine aquacole. Dans ce dernier cas, la qualité du milieu est vitale pour les espèces élevées et ou cultivées. De nombreuses données sont généralement acquises par les producteurs. Elles sont collectées manuellement ou au travers de capteurs, pour des systèmes de production. Il peut s'agir de la concentration en oxygène dissous, de la salinité, de la turbidité, de la température de l'eau, de la biomasse ainsi que de la quantité d'intrants et d'énergie dépensée.

2.4.1 L'aquaculture et la science de données

L'aquaculture est une industrie qui possède l'une des croissances les plus fortes dans le domaine de la production alimentaire à l'échelle mondiale [46].

Elle a connu une croissance majeure et à deux chiffres, durant cette dernière décennie. Tout en reconnaissant le potentiel que représente l'aquaculture pour le bien-être humain, cette activité a des impacts sociaux, économiques et environnementaux négatifs liés à

un certains types de pratiques. Certaines pratiques se sont révélées être non durables [61] alors que d'autres, malgré des résultats plus ou moins contrastés, en fonction des années, restent viables. Les innovations technologiques ont été et seront nécessaires pour une gestion plus durable de cette ressource.

D'après [100, 80], les systèmes intelligents intégrant l'apprentissage automatique à partir des données et permettant la gestion automatique des processus en aquaculture devraient permettre dans le futur une meilleure compréhension du marché, une réduction des maladies, principal frein à la durabilité de ce secteur, et une meilleure gestion de l'eau et de l'énergie dont certaines pratiques sont fortement consommatrices. Dans le cadre de la mise en place du processus complet d'extraction de connaissances, décrit en introduction, nous intéressons aux étapes qui le composent. Pour rappel, l'un des verrous scientifiques est d'intégrer dans ce processus, une méthodologie prenant en considération l'ensemble des données générées dans la filière; Par rapport au domaine aquacole, les étapes suivantes, relevées dans [186], sont mises en place :

1. **L'acquisition de données par** :
   (a) des caméras étanches permettant de capturer à différents instants, des images sous l'eau, d'espèces élevées.
   (b) des capteurs permettant d'obtenir régulièrement des données physico-chimiques sur la qualité de l'eau comme la température, la salinité, l'oxygène dissous, la turbidité...
   (c) des stations météorologiques permettant d'obtenir des données environnementales (pluviométrie, température ambiantes..).
   (d) des relevés réalisés manuellement par les éleveurs comme les paramètres physico-chimiques et zoo ou phytotechniques.
   (e) des laboratoires pour réaliser différentes analyses environnementales ou de santé des espèces élevées et /ou cultivés.

2. **Des systèmes de stockage de données** :
   (a) dans des bases de données relationnelles ou non.
   (b) dans des formulaires numériques ou papiers, répertoriés sous différents formats (xls,xml...).

3. **Des modèles d'apprentissages automatiques appliqués pour répondre aux problématiques et aux enjeux métiers** :

(a) des méthodes permettant d'extraire de nouvelles connaissances, de générer de nouveaux descripteurs qui serviront de données d'apprentissage

(b) des modèles d'apprentissages pour la détection de maladies à partir d'images, la modélisation de la biomasse dans le temps, la prédiction des paramètres physico-chimiques concernant la qualité des milieux de culture et/ou d'élevage....

2.4.1.1 Des outils d'aide à la décision

: Avant les années 2000, plusieurs logiciels destinés au contrôle des processus en aquaculture intégrant des méthodes en intelligence artificielle avaient déjà été développés. L'objectif de ces outils étaient de gérer différentes tâches telles que la collecte de données, l'analyse des tendances, et le contrôle automatique de la qualité du milieu, de la biomasse et de l'alimentation.

[43] présente un système de surveillance et de contrôle de l'environnement comprenant des capteurs essentiellement pour le suivi de la qualité de l'eau dont l'oxygène dissous, la salinité, la turbidité, la température et la biomasse. Mais à noter que la plupart des systèmes de contrôles ce sont intéressés à l'évolution de ces variables [99, 141].

[100] fait un état de l'art des applications intégrant des méthodes en IA pour le suivi de la qualité de l'eau principalement par l'utilisation de méthodes de la logique floue. Ces méthodes sont utilisées car les données contiennent une part d'imprécision non négligeable. En effet, dans ce dernier papier, l'acquisition des données est réalisée sur la base d'échantillons et les relevées sont réalisées dans des bassins tests, de petites tailles (de l'ordre du m²).

Pour des installations plus conséquentes, différents progiciels de systèmes de décision ont été développés par le biais de modules d'application intégrant des systèmes de décision et de normalisation des données. Concernant le suivi des paramètres physico-chimiques, l'application *AQUASMART* ([1]), dédiée à l'aquaculture a été développée afin d'améliorer les stratégies d'alimentation. Cet outil utilise un modèle de régression du taux de conversion alimentaire, ayant pour variables explicatives le poids moyen initial des poissons et la température moyenne.

2.4.2 Les objectifs les plus répandues associés aux méthodes en science de données

Analyse de l'impact de l'environnement d'élevage sur la productivité des systèmes.

Selon [100, 76], les aquaculteurs réalisent qu'en contrôlant les conditions environnementales et les intrants dans le système (ex. l'eau, l'oxygène, la température, le taux d'alimentation et la densité de peuplement), il est possible de limiter les facteurs de stress des espèces et les externalités (quantité d'effluents rejetés vers l'environnement littoral...).

Dans [13] les auteurs énoncent un modèle proche de la classification par paire, afin de pondérer l'influence des données environnementales (température, pH, oxygène dissous..) et l'alimentation en prenant comme cible la survie. Des relations importantes entre, la nourriture et la température de l'eau sur la survie ont été relevées. La survie des organismes est d'une importance capitale dans ces systèmes. Les résultats économiques dépendent très fortement de cette variable. Quelques tests ont mis en avant un lien potentiel de l'oxygène dissous sur la survie. La précision des modèles était supérieure à 90%. Cependant, dans la pluspart des papiers ([13, 76, 77]), les expérimentations ont été réalisées sur un nombre d'élevages très réduit, et sur de courtes périodes (et réalisées souvent dans des mésocosmes dans le cadre d'expérimentations).

Notons que de manière générale, les premières études utilisaient couramment la logique floue, appliquée à des données environnementales imprécises. En effet, comme énoncé, cette méthode permet d'apprendre avec des connaissances empiriques, décrites par exemple par des données catégorielles et souvent non explicites comme la quantité d'eau renouvelée (peu, beaucoup, etc.). Dans des papiers plus récents, l'étude de la teneur en oxygène dissous permet de travailler sur un indicateur intégral de la qualité de l'eau, car elle est utilisée pour évaluer l'état de santé des écosystèmes aquatiques et la capacité à maintenir les organismes aquatiques [155]. Une teneur insuffisante en oxygène dissous peut entraîner un stress, voir la mort des organismes aquatiques lorsque qu'il est consommé plus rapidement qu'il n'est produit (mise en place d'un système hétérotrophe). Par conséquent, la surveillance et la mesure de l'oxygène dissous est d'une importance cruciale dans les écosystèmes aquatiques pour la survie des espèces telles que les crevettes, les écrevisses et les poissons [92] Les études sur des données statiques agrégeant les données temporelles à des valeurs moyennes journalières par exemple, ou encore des études statistiques sur l'analyse des moyennes de deux échantillons on permis d'identifier certains seuils critiques. Concernant l'analyse de séries temporelles de l'oxygène dissous, le *RNN* présente des avantages évidents et est largement utilisé dans les systèmes de surveillance de cette variable. Sur la base du *RNN*, [185] ont adopté la technologie *k-PCA* pour réduire le bruit des données brutes et améliorer la précision. De meilleures performances de prédiction que les autres modèles, allant jusqu'à plus de 90% de précision. [22] ont amélioré le *RNN* et effectué un regroupement *K-means* sur les séries chronologiques d'oxygène dissous, améliorant ainsi la précision de la prédiction. [69] ont combiné des informations sur la qualité de l'eau avec des informations météorologiques. Ils ont appliqué la méthode de descente de gradient pour sélectionner les facteurs qui ont une plus grande influence sur l'oxygène dissous *OD*, et ont prédit cette variable par le modèle *LSTM*. Les résultats ont révélé que le modèle avait une bonne capacité de prédiction et de généralisation. Il a permit de déterminer un ordre d'influence entre plusieurs variables physico-chimiques et l'*OD*. Toutefois, l'intégration de plusieurs variables temporelles dans le domaine aquacole, n'est pas

courant dans la littérature.

La teneur en *OD* est aussi affectée par d'autres facteurs de qualité de l'eau, tel que la valeur de température [41].

2.4.2.0.1 Les études sur la biomasse : La gestion des ressources dans les systèmes de production aquacole mais aussi en halieutique, passe par la gestion de la biomasse des organismes. La biomasse est un paramètre important qui peut influencer la productivité des systèmes. Différents travaux décrits ci-dessous ont été menés afin d'estimer la taille et le poids des organismes, par des modèles d'apprentissages artificiels [117]. ( ) La littérature relève l'amélioration par les méthodes d'IA, de la qualité des données au moment de leur acquisition, en vue de leurs pré-traitements et d'une analyse avec des résultats satisfaisants.

Il existe différentes approches, plus traditionnelles, d'acquisition de données concernant l'estimation de la biomasse dans le domaine des pêches. Parmi ces méthodes, certaines sont basées sur la quantification du matériel utilisé pour la pêche (engins, filets..) et d'autres informations sur les prises réalisées (compositions des espèces capturées par taille et par âge). Ces informations permettent de définir des mesures relatives au taux de capture [112].

L'apprentissage automatique permet une estimation plus précise de la biomasse (la taille, le poids...), sans passer par des méthodes traditionnelles. [186]. Concernant l'estimation de la taille des espèces, les réseaux de neurones sont très souvent utilisés sur des données issues d'images. Différents types de données ont été utilisé, dans des recherches qui ont été menées pour la plupart sur des jeux de données accessibles en ligne (i.e *ImageNet*, jeu de données des poissons de l'Atlantique [35]). [120] ont proposé le modèle *R-CNN* sous différentes architectures pour l'estimation de la longueur du bar européen.

Néanmoins la complexité des images complique le pré-traitement des données; c'est le cas par exemple de la détection de poissons lors d'une situation de chevauchement sur l'image. Pour cela, [52] ont employé le modèle *Mask R-CNN* ([66]) et la technique du gradient local pour estimer la longueur des poissons dans l'Atlantique Nord, pour obtenir une segmentation précise.

[108] ont employé un modèle *CNN* fondé sur la méthode d'apprentissage par transfert pour estimer la longueur du poisson pour les données issues d'un petit échantillon prélevé dans un étang. Les résultats ont révélé que le modèle avait un bonne estimation et pouvait atteindre une précision de plus de 93 %.

En ce qui concerne l'estimation du poids de ces organismes, les chercheurs font en grande partie leurs prédictions en fonction des caractéristiques de la forme du corps des poissons, en utilisant des méthodes de traitement d'image pour extraire la taille du

poisson, la géométrie du dos et la surface du corps. [48] ont employé une combinaison de régression linéaire et *CNN* pour prédire le poids en divisant la surface du corps des poissons, pour essayer d'atteindre une précision de prédiction élevée. Les modèles restent limités à l'acquisition d'images à une distance fixe et à l'estimation automatique du poids.

Des modèles de production ont été les premiers à être utilisés dans l'analyse sur l'évolution de populations biologiques [142]. Ces modèles sont basés sur des modèles mathématiques pour l'estimation de taux de production, en considérant les pêches et d'autres facteurs naturelles (mortalité...).

La classification des espèces : L'apprentissage automatique, comme énoncé, permet d'extraire des caractéristiques dans les images. L'enjeu dans l'aquaculture est d'en extraire en vue d'une classification des espèces. Les développements de l'apprentissage automatique ont permis de réaliser des modèles de classification des espèces sur la base des caractéristiques des poissons. La performance des méthodes d'apprentissage est liée à la qualité des données, et donc à la description des entités. Les descripteurs doivent être identifiés pertinemment pour assurer une précision optimale. La recherche concernant la reconnaissance des espèces s'oriente sur l'extraction des caractéristiques les plus pertinentes.

Dans [78, 75], un pré-entraînement sur les ensembles de données de classification *ImageNet* ([35]) a été réalisé, pour acquérir les paramètres du modèle. Ensuite le modèle *CNN* a été optimisé grâce à l'ensemble de données réelles pour classer les poissons. [85] ont proposé une méthode pour l'extraction de caractéristiques d'arrière-plan, par une segmentation avec *Kmeans*. Ils ont aussi proposé une sélection et une extraction de caractéristiques à partir de la teinte, de la saturation et de la valeur des pixels, par une méthode de *feature selection*. Cette approche a permis d'extraire les caractéristiques (couleurs) des poissons *koï*. Ils ont ensuite utilisé le *SVM* pour une classification avec une précision de plus de 95 %.

[163] ont proposé une méthode de descripteur local pour extraire les caractéristiques de texture et de couleur. Ils ont utilisé l'Analyse Discriminante Linéaire ([171]) pour réduire le nombre d'éléments afin de distinguer les catégories. Le classificateur *AdaBoost* est ensuite employé pour classer les poissons avec une précision de plus de 96 %.

[31] ont utilisé les *SVM* et *KNN* pour la classification. Les résultats ont montré que la précision était supérieure à celle du modèle *CNN*.

[24] ont également proposé un modèle hybride d'apprentissage profond pour la détection et la classification des poissons, à partir d'un ensemble de données créé par eux-mêmes.

Ils existe néanmoins diverses contraintes liées pour la plupart aux mauvaises con-

ditions sous-marines (turbidité, manque d'éclairage). Il est aujourd'hui complexe dans ce cas de déterminer un modèle robuste à ces contraintes, et notamment lorsque la résolution du pixel est basse. Généralement, les méthodes de super-résolution d'une seule image ([176]), processus qui consiste à améliorer la résolution spatiale, permettent d'améliorer la précision de la classification des espèces de poissons.

[26] ont étudié un réseau de détection de poissons légers avec l'architecture d'apprentissage la plus profonde pour l'identification et la classification des poissons en réponse à des conditions difficiles, obtenant de bons résultats dans des eaux turbides.

Pour le problème des images à basse résolution collectées dans des conditions extrêmes, [160] et [130] ont proposé une méthode d'amélioration de la résolution pour parvenir à la classification des espèces.

Différents types de réseaux connexionnistes sont souvent entraînés avec des jeux de données en ligne. [135] et [149] ont développé une méthode combinant *CNN*, *SVM* et *KNN* pour classer les poissons, avec une précision de plus de 90%. A partir du jeu de données *Fish4Knowledge* ([51]) [131] ont proposé un modèle de réseaux de neurones pour la reconnaissance de poissons en eau profonde. Ils ont ensuite utilisé le *SVM* pour classer les poissons et ont obtenu un taux de précision de plus de 98 %.

Les enjeux dans le domaine aquacole, précédemment cités, sont les enjeux les plus important auxquels la science de données a jusqu'ici essayé de répondre. Les méthodes d'apprentissages automatiques ont été moins appliquées aux enjeux présentés par la suite.

2.4.3 Les objectifs les moins répandus

- **Les études sur la croissance** : Dans l'aquaculture très peu de techniques en machine learning ont été utilisées sur des données de croissance (taux de croissance...) dans le domaine des productions animales. Les modèles utilisés sont davantage basés sur des modèles mathématiques [19]. [182] évalue le potentiel des réseaux de neurones en tant qu'alternative aux fonctions de croissance qui ont été comparées dans [164]. On citera enfin l'utilisation de méthodes de *feature sélection* pour l'étude de la croissance des algues afin d'évaluer les facteurs favorisant leur expansion [136]. Les principaux facteurs sont liés à la qualité de l'eau, comme la température.

- **La génétique** : avec le développement de la bio-informatique, [60] classifient des espèces de poisson à partir de marqueurs génétiques. Ils utilisent des motifs générés par l'algorithme génétique. [56] utilisent le modèle du réseau neuronal Self-Organizing Map (*SOM*) pour détecter des groupes de structures génétiques.

- **L'impact environnemental de l'aquaculture** : [50] se sont concentrés sur l'effet environnementale et économique par une méthode naturelle de simulation de

différentes cultures ciblant une bonne performance.

- **L'évolution dynamique de l'activité des bassins** Grâce à la télédétection couplée aux méthodes de fouille de données (règles de classification, arbre de décision), [62] ont proposé une démarche de suivi dynamique des évolutions de l'activité des bassins à l'aide d'une série temporelles d'images satellitaires. [23, 187] ont repris ces travaux et proposent une démarche novatrice basée sur l'extraction de motifs dans un graphe dynamique attribué appliquée à la série temporelle d'images satellitaires, pour comprendre et suivre ces évolutions dynamiques.

2.5 Conclusion

Ce chapitre met en évidence la complexité des données issues des filières agricoles et aquacoles. Les objectifs et enjeux traités par les méthodes en science de données sont considérables. Cependant il y a encore des enjeux important non traités. Nous retiendrons l'intérêt de croiser des données de qualité du milieu, des caractéristiques zootechniques des espèces, et des paramètres édaphique en agriculture, comme en aquaculture, qui sont des données nécessaires à une meilleure gestion des systèmes de production. La dimension temporelle des données reste cependant peu prise en compte, alors que des méthodes de classification supervisées ou non supervisées à partir de données temporelles ont été appliquées dans le domaine agricole. Dans les deux cas, le clustering et la classification de séries temporelles multi-variées et multi-échelles, dont les définitions sont données ensuite, restent à notre connaissance peu ou pas pris en compte. Ces données temporelles sont régulièrement agrégées en données statiques (moyenne, minimum, maximum...) limitant ainsi l'information extraite.

Enfin, il n'existe pas de démarche intégrative permettant une analyse globale du fonctionnement et des résultats de ces filières, alors que de nombreux facteurs qu'ils soient techniques, environnementaux, sociaux, économiques sont à l'origine des performances de production. L'application des méthodes en science de données sur les données aquacoles, pourrait servir à mettre en place une méthodologie générale pour établir un lien entre pratiques et performances des élevages avec une vision mécanistique et pas uniquement statistique. Cette thèse propose une méthodologie et de nouveaux algorithmes pour répondre à ces enjeux.

Chapitre 3

Contribution méthodologique en sciences des données

Ce chapitre présente l'ensemble des contributions algorithmiques apportées durant la thèse pour l'analyse de séries temporelles en général et nous verrons, dans le chapitre 7, leur intérêt pour les données générées par la filière aquacole. De nouveaux algorithmes de clustering et de classification de séries temporelles ont été développés durant cette thèse. De plus, un nouveau formalisme de clustering de séries temporelles multi-variées est proposé. Des critères du clustering de séries temporelles multi-variées et multi-échelles seront proposés et viseront à regrouper des individus en tenant compte des variables temporelles communes. Les variables peuvent avoir des échelles temporelles différentes.
L'analyse des données environnementales du domaine aquacole par les nouvelles méthodes proposées, et qui sera présentée et commentée dans les chapitres suivants, montrera que l'amplitude des séries de données de qualité du milieu impacte considérablement les performances des espèces élevées. Le critère, que nous proposerons, sera pris en compte par les nouvelles méthodes. Avant d'analyser les données de la filière aquacole, les nouveaux algorithmes pour le clustering de séries temporelles multi-échelles, mono-variées et multi-variées, seront dans un premier temps détaillés et testés sur des jeux de données disponibles en ligne. L'objectif est de valider leur intérêt en comparant leurs performances avec d'autres méthodes existantes, telles que la méthode *K-Shape* [**?**] ou encore *KmeansTS* [71] qui seront présentées dans ce chapitre.

L'analyse de séries chronologiques est appliquée dans de nombreux domaines de l'ingénierie, des affaires, dans les finances, l'économie, la santé... Cette approche vise à extraire des connaissances utiles pour les experts à partir d'ensembles de données temporelles et massives ([8]). Elle est appliquée pour des objectifs variés tels que la correspondance de sous-séquences, la détection d'anomalies, la découverte de motifs [10], le regroupement, la classification, la visualisation, la segmentation et les prévisions. Notre étude porte sur le clustering de séries chronologiques.

Il existe 3 principales approches pour le clustering d'un ensemble de séries temporelles. La première approche est basée sur la construction de nouvelles caractéristiques :

les séries sont décrites par un vecteur d'attributs qui correspondent à des caractéristiques extraites ([82]). Afin de regrouper les individus, une méthode de regroupement classique (*K-Means*, *X-Means* [129], *DBscan*, ...) est appliquée sur ces caractéristiques. La seconde approche applique directement des méthodes de clustering existantes (essentiellement des approches k-Means) en utilisant des mesures de distance adaptées aux séries temporelles pour la comparaison de séries chronologiques. La troisième approche, détaillée ensuite, est basée sur le modèle en lui même. Au cours de la dernière décennie, les mesures de deux séries chronologiques ont été développées et améliorées en terme de performance. On trouve, par exemple, la mesure DTW (Dynamic Time Warping) [30], la mesure SBD (Shape Based Distance) [128] une amélioration de la corrélation croisée entre deux signaux. Nous nous baserons principalement sur ces deux mesures, notamment sur un ensemble de mesures dérivées de la mesure DTW qui est la mesure la plus adaptée et la plus utilisée pour le clustering de séries temporelles [39, 30], prenant en compte par exemple l'évolution des formes d'amplitudes dans le temps. Pour l'analyse de données dans le cadre de notre travail, nous proposerons des méthodes robustes intégrant la problématique de déphasage des séries par rapport aux axes (abscisse et ordonné).

La qualité d'un clustering dépend fortement de la mesure de distance et la manière dont les individus sont pris en compte par cette mesure. Le choix de la mesure de distance peut dépendre également du domaine, et en particulier des invariances requises par le domaine ([36]). Par exemple, les données de capture de mouvement nécessitent généralement l'invariance de la "déformation" (accélérations locales non linéaires) [91], et les données sur l'expression des gènes nécessitent généralement une invariance pour une mise à l'échelle uniforme ("étirement" linéaire) [89]. Plusieurs techniques conçues pour mesurer efficacement la similitude entre les séries chronologiques avec l'invariance ont été développées durant la dernière décennie.

Dans certains domaines, le regroupement des séries temporelles doit être fait en considérant davantage l'invariance et l'intervalle des mesures sur l'axe des y tout en considérant le décalage des séries sur l'axe des x. En effet, bien que des séries soient similaires en termes d'évolution de forme, elles peuvent être considérées appartenant à des classes différentes si, sur l'axe des y, leurs valeurs varient dans des intervalles distants. Selon le domaine, l'intervalle des valeurs sur l'axe des y peut discriminer fortement les classes d'appartenance. Par exemple, dans les domaines agricoles ou aquacoles, l'intervalle des valeurs en y des séries temporelles liées aux données environnementales, comme l'évolution de la variabilité de la température au-delà d'un certain degré, influe considérablement sur la croissance et la survie des espèces vivantes. Les méthodes de clustering de séries temporelles existantes ne considèrent pas directement l'intervalle des valeurs autour duquel les individus de chaque cluster se trouvent.

Les nouvelles approches proposées dans ce chapitre, s'appuient sur des méthodes

existantes qui prennent en compte le décalage des séries sur l'axe des x. Prenons comme exemple les séries de la figure 3.1. Malgré le fait que ces séries soient similaires en termes d'évolution de forme, elles peuvent être considérées dans des classes différentes si, sur l'axe des y, leurs valeurs varient dans des intervalles distants.

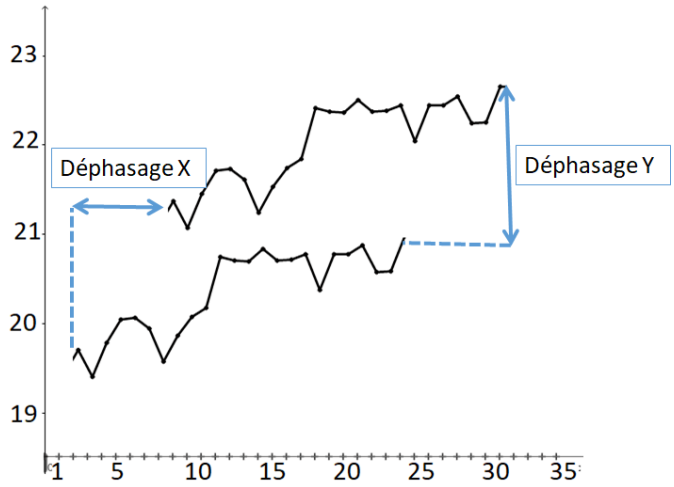

Fig. 3.1 Exemple de deux séries avec un déphasage sur les axes X et Y

Pour obtenir des clusters de séries temporelles dont les variabilités d'amplitudes se trouvent dans un intervalle restreint, il est nécessaire par exemple avec les méthodes existantes, de paramétrer un nombre de clusters important. En effet, il n'existe pas de méthode permettant d'obtenir automatiquement un nombre de clusters de séries temporelles en fonction de l'intervalle dans lequel évoluent les individus.

Dans ce chapitre, nous proposerons d'abord une nouvelle méthode de clustering des séries temporelles mono-variées, s'appuyant sur une approche basée sur l'analyse des formes et prenant en compte la variance selon l'axe des y. D'autre part, la stratégie de la méthode que nous proposons permet de définir automatiquement un nombre k optimal de clusters, en utilisant un critère de dispersion à l'intérieur d'un cluster. Contrairement à la plupart des méthodes qui normalisent les données, notre approche s'applique tant sur les séries temporelles normalisées que les séries temporelles brutes. Cette nouvelle méthode est robuste au décalage des séries sur l'axe des x car on utilise des métriques qui prennent en compte la déformation des séries dans le temps. Pour le décalage en y, on considérera un intervalle maximal sur lequel ces mesures varient. Pour cela nous considérons l'étalement des distances entre les séries et leur représentant intra-cluster.

3.1 Les mesures de dispersions

Les mesures de dispersions servent à caractériser l'étalement des valeurs présentes dans une distribution. Plus la distribution sera étalée, plus la valeur de la mesure de dispersion sera élevée. Ce critère de dispersion s'applique pour toutes les mesures suivantes :

- L'étendue [55] : quantifie la longueur de l'intervalle dans lequel se situe les valeurs de la distribution;
- L'écart moyen [181]: la moyenne des écarts à la moyenne des valeurs de la distribution.
- La variance [88] est la moyenne du carré des écarts à la moyenne des valeurs de la distribution.
- L'écart type [12], la racine carrée de la variance, mesure la dispersion des valeurs d'un échantillon statistique ou d'une distribution de probabilité.

Nous avons vu dans le chapitre précédent que certaines méthodes comme *K-Means* vise à identifier la valeur la plus représentative de chaque cluster, comme le centre des distributions. Pour décrire l'ensemble des données, il faut également mesurer l'étalement des valeurs autour du centre. Pour cela, nous proposons une nouvelle mesure de dispersion des mesures de distance utilisées pour le regroupement des séries temporelles, et notamment de la distribution intra-cluster entre les individus et le représentant (centroïde). Celle nouvelle mesure de dispersion fait intervenir la mesure d'entropie qui détermine la quantité d'information. Elle est aussi décrit dans la littérature comme une mesure de désorganisation car elle augmente avec l'équilibre des probabilités des valeurs dans une distributions.

L'origine de ces recherches sur la théorie de l'information remonte aux études entreprises en physique et en mathématique par *Boltzmann* et *Markov* [151, 115] sur la notion de probabilité d'un événement et les possibilités de mesure de cette probabilité. Le modèle permet de modéliser la dynamique des séries temporelles. *Shannon* donne à la notion d'information un statut physique à part entière [29]. Il considère pour cela que l'information est liée à la "redondance" ou au "bruit", dans le cadre de la transmission de données par des canaux de communication entre une source et un récepteur (capteur). Si l'on considère un surplus d'information inutile lors de la transmission alors l'information peut contenir des données redondantes ou encore du bruit. La théorie de l'information minimale représente un critère de la théorie de l'information et est une estimation d'une mesure d'ajustement du modèle de clustering. [132] propose un test pour la qualité de l'ajustement des modèles auto-régressifs (AR). L'idée de ce test de Quenouille a été étendue à un test de qualité d'ajustement des modèles de moyenne mobile (MA) [72].

Nous verrons comment la nouvelle mesure de dispersion, qui considère la quantité d'informations par la mesure d'entropie, s'appliquera aux approches de clustering existantes. Le regroupement de ces données peut être complexe car les distributions des distances entre les individus et leur représentant ont, intra-cluster, des probabilités

déséquilibrés et peuvent avoir une étendue importante. Notre deuxième contribution concerne l'extension de notre méthode pour un clustering des séries temporelles multivariées.

3.2 Des mesures de distance adaptées aux séries temporelles

Plusieurs mesures de distance (ou de similarité) ont été proposées pour calculer une distance entre deux séries temporelles. Dans beaucoup d'approches de clustering, elles peuvent influer sur la sélection des individus par cluster. La mesure appelée Shape Based Distance (*SBD*) utilisée dans la méthode de clustering dans [128], est basée sur la corrélation croisée de deux signaux avec un calcul efficace de la transformée de Fourier [27].

Une formalisation d'une mesure de similarité non métrique, proposée par [173], appelée *LCSS*, est basée sur le calcul de la plus longue sous-séquence commune entre deux séries et attribue plus de poids aux parties similaires de ces deux séries.

En général, les distances entre séries temporelles développées rentrent dans deux catégories de distances : celles qui sont des *LP-normes* et celles qui ne sont pas métriques. La première catégorie de distances ne supporte pas le décalage temporel; en revanche la deuxième catégorie gère très bien ce décalage. [20] propose de croiser une distance métrique (*L1-normée*) et la distance d'édition pour construire une distance métrique supportant le décalage temporel local. Cette distance appelée "Edit distance with Real Penality" (ERP) modifie la distance d'édition en ajoutant une pénalité réelle liée au nombre d'opérations nécessaires pour que deux séquences soient similaires. Toutes ces mesures de distance comparent des séries en considérant uniquement les effets du déphasage temporel. Elles ont été combinées avec des approches de clustering de base (*kmeans*), qui ne prennent pas en compte les dérives dites d'amplitude pour le regroupement de séries temporelles. [57] présentent un état de l'art complet recensant les mesures développées. Nous trouverons davantage des mesures robustes au décalage des séries sur l'axe des x. Les mesures que nous verrons plus loin et qui ont été davantage exploitées durant cette thèse, sont principalement des dérivées de la célèbre distance de distorsion temporelle *DTW* ([124]).

La distorsion temporelle *DTW* ([124]) est une distance calculée à partir d'un chemin (entre les premiers et les derniers points de deux séries) présentant un alignement non linéaire optimal entre deux séries temporelles. Soient $q = \{q(1), \ldots, q(n)\}$ et $c = \{c(1), \ldots, c(n)\}$ deux séries temporelles de la même variable mesurée. Le calcul de la distance *DTW* consiste dans un premier temps à construire une matrice M de dimension $n \times n$ où $M(i,j) = (q(i) - c(j))^2$, comme le montre la figure 3.2.

On appelle chemin déformé $W = \{w_1, \ldots, w_r, \ldots, w_p\}$ où $p \geq n$, une suite d'éléments de la matrice M qui sont contigus et tels que le premier élément est $w_1 =$

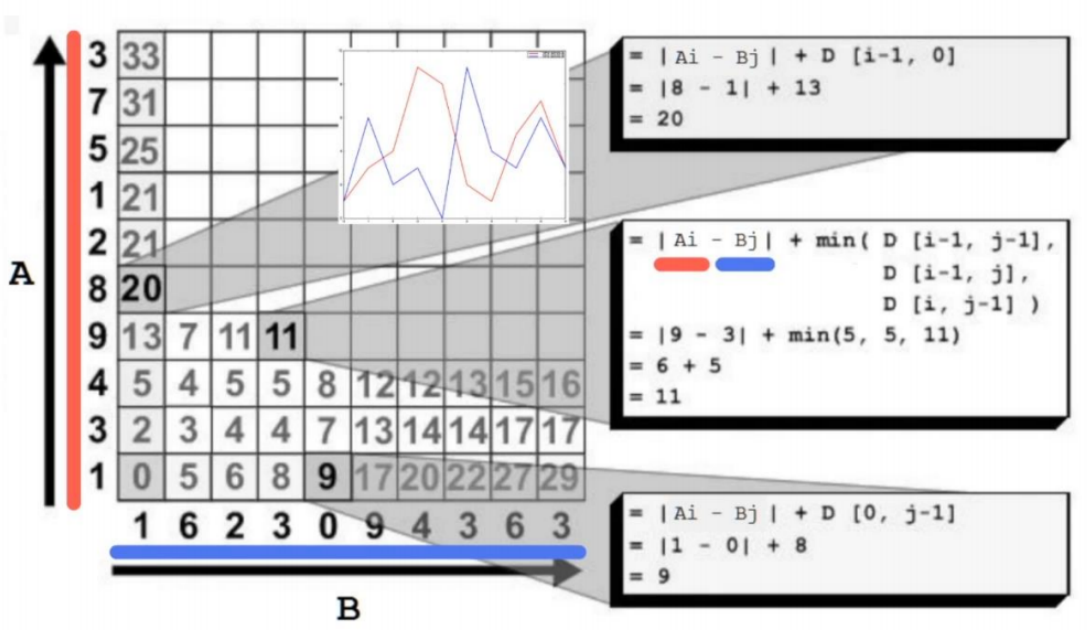

Fig. 3.2 Construction de la matrice *DTW*, image extraite de [96]

$M(1,1)$ et le dernier est $w_p = M(n,n)$ et $w_r = M(i,j)$ avec i et $j \in [0, n-1]$. La méthode détermine l'alignement optimal en recherchant dans la matrice le chemin W qui minimise la distance (euclidienne) cumulée i.e $Wo = argmin(\sqrt{\sum_{i=1}^{p} w_i})$.

La figure 3.3 détaille la recherche de l'alignement optimal en commençant à partir du dernier point calculé de la matrice M. C'est-à-dire qu'à partir du point (M (1, n)), la recherche de l'alignement se fait de manière itérative en sélectionnant la cellule de valeur minimale dans le voisinage contenant les cellules avec les coordonnées suivantes : *(i-1 , j), (i , j-1) (i-1 , j-1)* avec *i* la ligne et *j* la colonne de la cellule en cours. L'opération est répétée jusqu'à ce que la cellule M(n,1) soit sélectionnée.

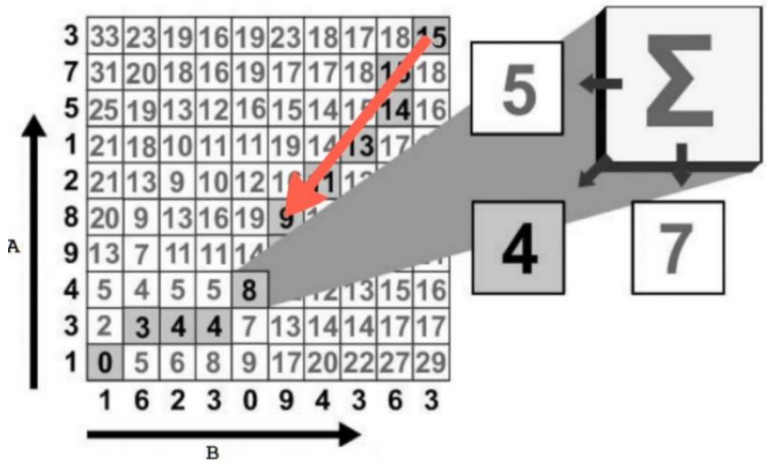

Fig. 3.3 Recherche optimale de la méthode *DTW*, image extraite de [96]

La recherche du chemin optimal est coûteuse, étant de complexité $O(n^2)$, différentes optimisations de la mesure *DTW* existent. Les principales optimisations visent à réduire l'espace de recherche en introduisant des contraintes sur la zone de recherche. [148] utilisent comme zone de recherche, dans la matrice *M*, une bande autour de la diagonale inversée (anti-bande), [73] cherche le chemin optimal dans une zone définie par un parallélogramme autour de la diagonale inversée. La figure 3.4 présente, pour ces deux approches, la contrainte dans l'espace de recherche du chemin optimal entre la série en bleu et celle en rouge.

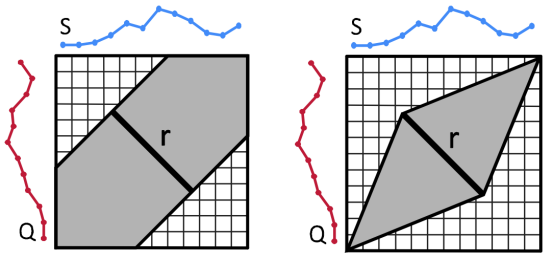

Fig. 3.4 Contrainte de *DTW* à l'aide de la : bande de Sakoe-Chiba (gauche) ; parallélogramme d'Itakura (droite) [53].

Les méthodes $FastDTW$ ([150]) et $multiscaleDTW$ ([38]) optimisent la recherche spatiale suivant une approche à plusieurs niveaux de résolution dans la matrice M. La figure 3.5 présente les trois étapes d'optimisation de l'approche $FastDTW$ qui consistent à :

1. réduire la résolution d'une série chronologique;
2. utiliser un chemin à faible résolution comme solution initiale à une résolution plus élevée;
3. affiner une trajectoire par un ajustement locale.

Le passage de la faible résolution à une résolution plus élevée peut se faire par dichotomie.

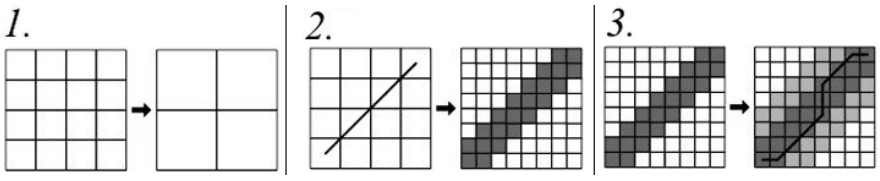

Fig. 3.5 Optimisation spatiale de *DTW* par l'approche $FastDTW$. [150]

[90] proposent une bande inférieure appelée bande de *Kheog*, qui a été couplée à la distance DTW. Une valeur est associée à la bande de *Kheog* (figure 3.6) et permet de déterminer en fonction d'un seuil, si une série est proche ou non d'un représentant.

$$LB\_Keogh(Q,C)=\sum_{i=1}^{n}\begin{cases}(q_i-U_i)^2 & if\ q_i>U_i\\ (q_i-L_i)^2 & if\ q_i<L_i\\ 0 & otherwise\end{cases}$$

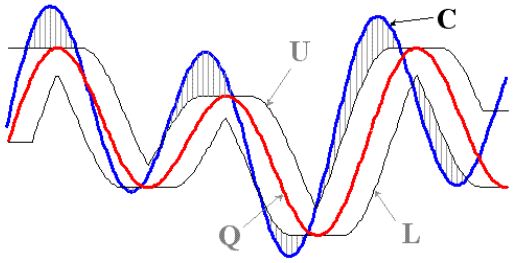

Fig. 3.6 Calcul de la bande de *Kheog* [90]

Les méthodes existantes ne s'intéressent pas à la quantité d'informations obtenue ou non, suite au clustering des individus. C'est à dire à l'équiprobabilité des valeurs de distances entre les individus et leur représentant. Elles s'intéressent d'avantage à la réduction de ces distances, en fonction de la mesure de distance utilisée et qui peut tenir compte par exemple de l'homogénéité des tendances des séries qui se trouvent dans un même cluster.

3.3 Les principales approches de clustering de séries temporelles

Pour rappel, on appelle k-clustering C de S, l'ensemble $C = \{C_1, C_2, \ldots, C_k\}$ contenant k sous-ensembles homogènes de S (au sens d'une mesure de distance $Dist$ adaptée aux séries temporelles), chaque cluster admet un représentant. Le représentant appelé également prototype, qui peut être un centroïde, sera noté R_{C_i} .
Un k-clustering $C = \{C_1, C_2, \ldots, C_k\}$ avec $\forall i \in \{1, \ldots, k\}$ et $C_i = \{i_1, i_2, \ldots, i_{m_i}\}$ ensemble de m_i individus représentés par leur série temporelle $\{s_{i_1}, s_{i_2}, \ldots, s_{i_{m_i}}\}$ devra vérifier les critères suivants :

1. $S = \cup_{i=1}^{k} C_i$ et $C_i \cap C_j = \emptyset\ \forall i \neq j$.
2. $Dist(R_{c_i}, s) < Dist(R_{c_j}, s)\ \forall s \in C_i$ et $j \neq i$.

On notera dans la littérature trois grandes approches de clustering de séries temporelles (3.7), la première est fondée sur l'extraction de caractéristiques, la seconde sur un modèle, et notamment le modèle de *Markov* [42] qui permet de représenter des séquences d'évènements et qui sera détaillé ensuite. Et la troisième approche, celle principalement utilisée dans cette thèse, est l'approche basée sur les formes, c'est à dire sur les variations des amplitudes des séries. Nous présenterons ensuite les principales

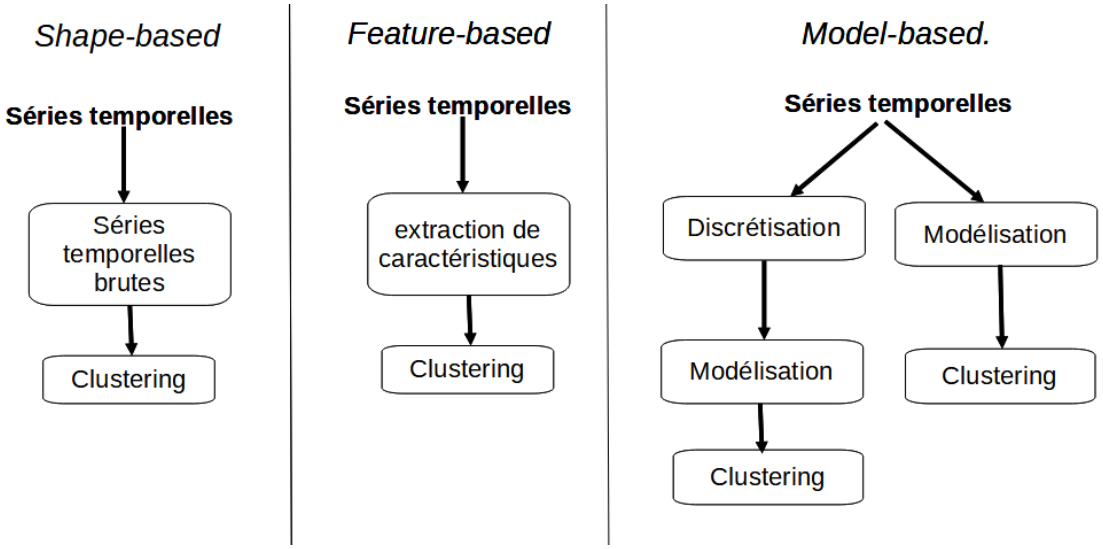

Fig. 3.7 Les grandes approches de clustering de séries temporelles. [2]

méthodes selon ces trois approches, nous intéresserons principalement à la présentation des méthodes basées sur les formes.

Dans la première approche de clustering de séries temporelles fondée sur l'extraction de caractéristiques, ces dernières représentent principalement des statistiques dérivées de chacune des séries, comme par exemple leur moyenne, leur écart type.... Ces caractéristiques servent à réduire la dimensionnalité des séries temporelles originales, lorsqu'elle est trop importante. Pour regrouper des séries temporelles, ces statistiques, peuvent être utilisées en entrée des méthodes de clustering standards (*K-means* [65], clustering hiérarchique [81], *SOM* [93] ...). L'avantage d'utiliser ces caractéristiques est qu'elles permettent de regrouper des séries temporelles de longueurs différentes.

3.3.1 Approche de clustering de séries temporelles basées sur les caractéristiques

Dans [58], les auteurs ont regroupé les séries temporelles d'imageries par résonance magnétique fonctionnelle (IRMf) en groupes de voxels présentant des activations similaires à l'aide de deux algorithmes : *k-means* et le regroupement hiérarchique de *Ward* [81]. Le voxel en plus de sa position spatiale, permet de stocker une information physique (couleur, intensité...). L'espace des caractéristiques (attributs), utilisé (dans [58]) par les méthodes de clustering (K-means et hiérarchique), était lié aux valeurs obtenues par la fonction de corrélation croisée, entre l'activation IRMf et le paradigme (ou stimulus). [59] ont également utilisé d'autres caractéristiques, comme le délai et la force d'activation qui sont mesurés par voxel, pour montrer qu'il était possible d'identifier les régions présentant des délais d'activations et des (taux d') activations significativement différents, suite à l'application de *k-means* sur ces caractéristiques. [177] ont modifié l'algorithme standard de clustering *K-means* pour la reconnaissance

de mots, représentés par un signal acoustique. Pour cette reconnaissance, des caractéristiques issues des signaux (des mots) ont été générés et utilisées, par la méthode *K-means* modifiée. Ces caractéristiques ont été obtenues par la méthode de quantification de vecteur, qui consiste à représenter un vecteur de dimension *k* par un autre vecteur de même dimension, mais ce dernier appartenant à un ensemble fini. La principale modification apportée à la méthode *K-means*, était la subdivision du cluster avec l'inertie la plus élevée; L'ensemble des points sont considérés dans un même cluster à l'état initial; Si les clusters générés ne converge pas avec un nombre *k* (initialisé à 1), alors k est incrémenté, et *K-means* est ré-appliqué (avec k=2). Un nombre maximum d'itération est fixé, en cas de non-convergence. A la fin, le cluster avec l'inertie la plus élevée, est re-partitionné (avec k=1) selon le même procédé. L'algorithme de clustering *k-means* modifié (*MKM*) proposé s'est révélé plus performant que l'algorithme de clustering couramment utilisé à l'époque : l'algorithme *UWA* (unsupervised without averaging) [134]. Ce dernier se concentre, sur le cluster qu'il génère avec le plus d'individus, de trouver tous les modèles de mots "proches" du centre de ce groupe, de les éliminer de l'ensemble et de regrouper à nouveau les modèles restants.

[174] ont présenté une approche permettant d'effectuer un regroupement de séries temporelles, sur diverses résolutions en utilisant la transformée en ondelettes de *Haar* ([159]). Cette transformée permet de décomposer un signal, en fonction de base et selon différentes résolutions. La décomposition en ondelettes de *Haar* est calculée pour toutes les séries temporelles. L'algorithme de clustering *k-means* est appliqué, en commençant par le niveau de résolution le plus élevé. A partir de ces clusters obtenus sur ce niveau, *k-means* est appliqué progressivement vers des niveaux de plus en plus fins.

3.3.2 Approche de clustering de séries temporelles basées sur le modèle

Cette approche considère que chaque série temporelle est générée par un certain type de modèle. Prenons le cas des modèles les plus exploités pour cette approche qui sont le modèle de *Markov* [42] et le modèle *ARMA* [25].

Le modèle de Markov : Supposons que nous observions une série temporelle $s = (s_1, s_2, s_3, ..., s_{i-1}, s_i, ..., s_n)$, où chaque s_i est l'un des états $1, ..., s$ d'une variable *S*. Le processus générant la séquence *s* est un modèle de Markov *MC* si la probabilité conditionnelle que la variable visite l'état *j* au temps *t*. La séquence $(s_1, s_2, ..., s_{t-1})$, est seulement une fonction de l'état visité au temps *t-1*. Par conséquent, nous écrivons $p(xt = j|(s_0, s_1, s_2, ..., s_{t-1})) = p(s_t = j|s_{t-1})$ pour tout s_t dans s. En d'autres termes, la distribution de probabilité de la variable s au temps t, dite S_t, est conditionnellement indépendante des valeurs $(s_0, s_1, s_2, ..., s_{t-2})$, une fois que s_{t-1} est connu. Cette hypothèse d'indépendance conditionnelle nous permet de représenter une *MC* comme un vecteur de probabilités $p_0 = (p_{01}, p_{02}, ..., p_{0s})$, dénotant la distribu-

tion de S_0 (l'état initial de la chaîne) et une matrice P de probabilités de transition, où $p_{ij} = p(S_t = j|S_{t-1} = i)$.

Les processus AR (AutoRegressive), MA et ARMA [25] : On dit qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un processus AR d'ordre (p) , noté $\mathrm{AR}(p)$, si : $\forall t \in \mathbb{Z} : X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$ où $(\varphi_1, \ldots, \varphi_p \in \mathbb{R}^p)$ et $(\varphi_p \neq 0)$. La modélisation de (X_t) se résume à une relation linéaire le liant aux (p) derniers instants.

On dit qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un processus MA (Moving Average) d'ordre (q) , noté $(\mathrm{MA}(q))$, si : $\forall t \in \mathbb{Z} : X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$ où $(\theta_1, \ldots, \theta_q \in \mathbb{R}^q)$ et $(\theta_q \neq 0)$. On considère ici que le processus est la résultante d'une combinaison linéaire de perturbations dé-corrélées (un bruit blanc et son passé).

Les processus ARMA est une synthèse des processus AR et MA : $X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$.

Dans le cas du clustering des séries selon une approche basée sur le modèle, les séries temporelles sont considérées comme similaires lorsque les modèles caractérisant les séries individuelles ou les résidus restants après ajustement du modèle sont similaires. [138] ont présenté *Bayesian clustering by dynamics* (*BCD*) : un algorithme bayésien pour le regroupement dynamique. Étant donné un ensemble *S* de *n* nombres de séries temporelles univariées à valeurs discrètes, *BCD* transforme chaque série en une chaîne de *Markov* (MC) et regroupe ensuite des *MC* similaires pour découvrir l'ensemble le plus probable de processus générateurs. [83] ont étudié le regroupement de séries temporelles par *ARIMA* [154], en utilisant la distance euclidienne entre des coefficients cepstraux de deux séries comme mesure de dissimilarité. Le cepstre d'un signal est une transformation de ce signal du domaine temporel vers un autre domaine analogue au domaine temporel. Dans [83] les coefficients cepstraux des séries temporelles sont dérivés des coefficients d'auto-régression (*AR*). [116] a développé une procédure de regroupement hiérarchique qui est une procédure de clustering qui se base sur la p-valeur d'un test d'hypothèse appliqué à chaque paire de séries temporelles stationnaires données. En supposant que chaque série temporelle stationnaire peut être ajustée par un modèle linéaire d'auto-régression désigné par un vecteur de paramètres π = [1, 2,..., k], Un test statistique pour tester l'hypothèse selon laquelle il n'y a pas de différence entre les processus de génération de deux séries temporelles stationnaires. Deux séries sont regroupées si la p-valeur associée est supérieure au niveau de signification fixée au préalable.

3.3.3 Approche de clustering de séries temporelles basées sur les formes

Dans l'approche basée sur la forme : les formes de deux séries temporelles sont appariées aussi bien que possible, par un étirement et une contraction non linéaire de

l'axe temporel. Cette approche travaille directement avec les données brutes des séries chronologiques. Les mesures de distance/similarité utilisées et adaptées pour comparer deux séries, sont celles présentées dans la section 3.2. Pour assurer le clustering des séries, les distances calculées par ces mesures adaptées sont utilisées par des méthodes de clustering standards, compatibles avec les données statiques (*Kmeans*...). Quelques méthodes répandues, et les plus performantes dans la catégories des approches basées sur les formes, seront présentées et notamment l'approche *TSK-means* [71], *K-Shape*, et la combinaison de *K-means* avec la distance *DTW* et ses dérivés. Notons néanmoins qu'il y a peu d'approches de clustering, basées sur les 'formes', et prenant en compte des séries de tailles différentes. L'algorithme *Globale Alignement Kernel* (GAK) [170], est une approche de clustering qui prend en compte cette différence. Elle est basée sur un type de noyau qui détermine le calcul des distances entre deux séries afin de générer la matrice de distances, comme celle générée par la mesure *DTW* (pour le calcul de la distance entre entre deux séries). Le type de noyau est paramétré en fonction de la distribution des données (Gaussienne...).

K-meansDTW, l'algorithme *K-means* couplé à *DTW* [5] : L'approche de clustering Kmeans a été couplé aux mesures de distances adaptées aux séries temporelles. Par cela, *DTW* est une mesure régulièrement utilisée. Le calcul des distances entre le représentant et les séries se font donc selon cette mesure d'après le pseudo code de l'algorithme 5. Cette approche *KmeansDTW* peut être utilisées avec l'ensemble des

Algorithm 5 KmeansDTW

Input: $S = \{s_1, s_2, \ldots, s_n\}$ ensemble de séries temporelles

1: choisir aléatoirement k centroïdes sur les séries en entrée S
2: **while** pas de convergence **do**
3: attribuez les séries au centroïde le plus proche, en fonction de la mesure *DTW*
4: Recalculez les centroïdes
5: **end while**

mesures dérivées de DTW (DTW Itakura...).

L'algorithme TSK-means [71] : Cette méthode utilise le même principe que l'approche *KmeansDTW*, mais elle diffère sur le calcul de la distance entre les séries qui sera détaillée dans ce paragraphe. Soit $S = \{S_1, S_2, \ldots, S_n\}$ un ensemble de *n* objets de séries temporelles. Chaque objet $S_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}$ est caractérisé par *m* valeurs par rapport à *m* estampilles de temps. La matrice d'appartenance *U* est une matrice binaire $n \times k$, où l'élément $u_{ip} = 1$ indique qu'un objet de série temporelle *i* est affecté au cluster *p* ; sinon, il ne l'est pas. Les centroïdes de *k* clusters sont représentés par un ensemble de *k* vecteurs $Z = \{Z_1, Z_2, \ldots, Z_n\}$. $W = \{W_1, W_2, \ldots, W_n\}$ est un ensemble représentant les poids des marques temporelles sur chaque cluster. La valeur de l'élément w_{pj} indique le poids du $j^{\text{ème}}$ estampille de temps pour le $p^{\text{ème}}$ cluster. La fonction objective de *TSK-means* est formulée comme suit :

$$\sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ip} * w_{pj}(s_{ij} - z_{pj})^2 + (1/2) * \alpha * \sum_{p=1}^{k}\sum_{j=1}^{m-1}(w_{ij} - w_{pj+1})^2 \tag{3.1}$$

sous réserve de :

$$\begin{array}{l} \sum_{p=1}^{k} u_{ip} = 1, u_{ip} \in 0, 1 \\ \sum_{j=1}^{k} w_{pj} = 1, 0 < w_{pj} < 1, \end{array} \tag{3.2}$$

où α est un paramètre qui est utilisé pour équilibrer les effets entre les dispersions d'objets au sein des clusters et le lissage des poids des estampilles de temps. Le lissage des poids entre les estampilles de temps adjacents augmente avec l'incrémentation de la valeur de α. Le premier élément de la fonction objectif Eq. 3.1 vise à minimiser la somme des dispersions de tous les clusters. Le deuxième élément de la fonction objectif Eq. 3.1 consiste à lisser les poids des estampilles adjacents. Dans le processus de clustering, cette fonction objectif minimise simultanément la dispersion au sein du cluster et lisse les poids des estampilles adjacents.

L'algorithme K-shape [127] : L'algorithme 7, présente la méthode *K-shape* [128]. On peut s'apercevoir que le principe de l'approche est comparable à celui de la méthode *K-meansDTW* [5]. Il y a en effet, un calcul de distances entre les séries et leur représentant, et une recherche de convergences de l'attribution des séries à un cluster. *K-shape* utilise cependant la mesure *Shape Based Distance* (*SBD*) ([128]).

Les méthodes que nous proposons, s'intéressent à la variance de l'axe des *x* et celle des amplitudes (i.e axe des *y*). Pour cela, la stratégie utilisée est basée sur la symétrie de la distribution, des distances entre les individus et leur représentant, de clusters qui sont générés par une méthode existante.

On remarquera dans la section suivante 3.4, que dans le cadre des approches de clustering basées sur les formes, la distribution des distances $Dist(C)$ intra-cluster (entre les individus et leur représentant), suit une loi normale. Pour cela une étude de cette distribution est réalisée dans cette section, à partir de clusters de séries de données environnementales. Ces données sont issues de la filière aquacole étudiée dans les chapitres 6 et 7.

3.4 Étude des clusters de séries temporelles en fonction de la distribution des individus

La figure 3.8 affichent les séries temporelles regroupées en 3 clusters par la méthode *K-Shape*. Les séries sont issues de la filière aquacole présentées et analysées plus en détail dans les chapitres suivants. Ces séries représentent l'évolution de la température de l'eau des élevages qui seront étudiés dans le chapitre 6.

Algorithm 6 K-Shape

Input:
Output:

```
iter = 0
IDX0 = [ ]
while IDX ! = IDX0 et iter ¡ 100 do
    IDX0 = IDX
    for j = 1to k do
        X0 = [ ]
        for i = 1à n do
            si IDX(i) = j alors
            X0 ← [X0;X(i)]
        end for
        C(j) = ShapeExtraction(X0,C(j))
    end for
    for i = 1 à n do
        mindist = ∞
        for j = 1 à k do
            [dist,x0] = SBD(C(j),X(i))
            if dist ¡ mindist then
                mindist = dist
                IDX(i) = j
            end if
        end for
    end for
    iter = iter + 1
end while
```

Algorithm 7 ShapeExtraction

Input:
- X est une matrice n par m avec des séries temporelles z-normalisées.

Output:

```
X0 ← [ ]
for i ← 1 to n do
    [dist,x0] = SBD(C,X(i)) // Shape Based Distance
    X' = [X' ;x' ]
end for
S = X'^T . X'
Q = I - (1/m) . O
M = Q^T.S.Q
C0 = Eig(M,1) //
```

Les représentants des clusters colorés en rouge sont les centroides. On remarque qu'avec k=3 la méthode regroupe les séries selon deux tendances majeures (croissante, décroissante, et une dernière tendance légèrement concave). Dans ces clusters les am-

plitudes minimale et maximale des séries varient fortement entre les individus;

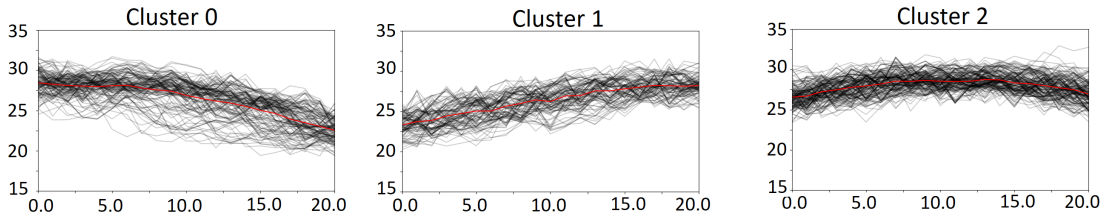

Fig. 3.8 Cluster par la méthode K-Shape avec k = 3, des séries temporelles de température

Il est possible de réduire cette variation, intra-cluster, avec les méthodes de clustering existantes, en augmentant le nombre de clusters souhaités. On peut voir dans la figure 3.9, 12 clusters générés par *K-shape*, sur les données d'évolution de température, où les séries sont plus proches de leurs représentants que les séries de la figure 3.8 (où k=3).

Néanmoins, l'opération qui consiste à déterminer un nombre optimal de clusters, selon l'amplitude des séries, reste une opération complexe. Pour arriver à obtenir ce nombre optimal, nous étudierons la dispersion des valeurs, intra-cluster, des distances entre les séries et leur représentant. Plusieurs tests ont été effectués avec d'autres variables afin de comparer les mêmes statistiques sur ces mesures. Le choix s'est porté sur des variables temporelles qui varient fortement au cours du temps. Ce qui est le cas pour la variable d'oxygène dissous (OD) étudiée ensuite, et qui, contrairement à la variable température, présente, des valeurs journalières avec un écart-type important.

La dispersion des valeurs *DTW*des données environnementales, et la mesure de dispersion : La figure 3.10 montre la dispersion des mesures *DTW*, intra-cluster, des 3 clusters de la figure 3.8, entre chaque instance et leur représentant. La figure 3.11 montre la dispersion des mesures de distances *DTW*, entre les individus et leur représentant, pour 2 clusters des séries temporelles d'oxygènes dissous sur une période de 20 semaines.
On observe à l'intérieur des clusters, une distribution normale (gaussienne ou gamma) des mesures de ces distances DTW. Cette distribution s'explique, pour rappel, par les algorithmes qui agrègent des individus autour du représentant (i.e la série moyenne).

Statistiques sur les distances entre les séries et leur représentant, par cluster de séries temporelles : La figure 3.12 présente les statistiques descriptives (moyenne, écart type, minimal, maximal) des distances *DTW* intra-cluster, pour le cluster *0* de séries temporelles de température obtenue précédemment par la méthode *K-shape*. Elle

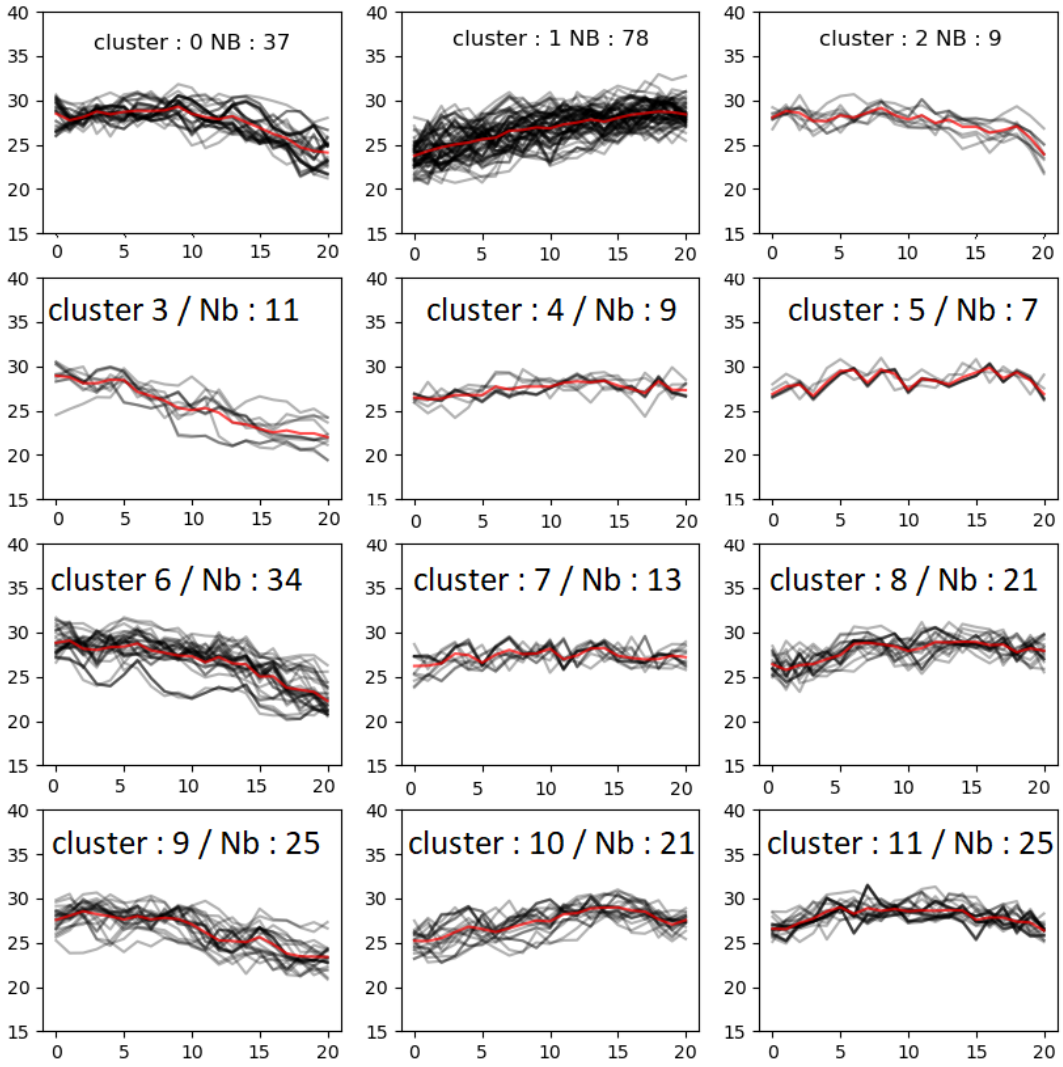

Fig. 3.9 Cluster de séries temporelles liées à la température par la méthode K-Shape avec k = 12

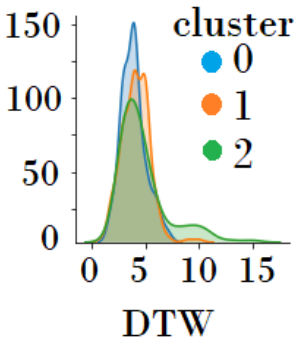

Fig. 3.10 Distribution des distances *DTW* entre les séries et leur représentant par cluster de température générés par K-Shape

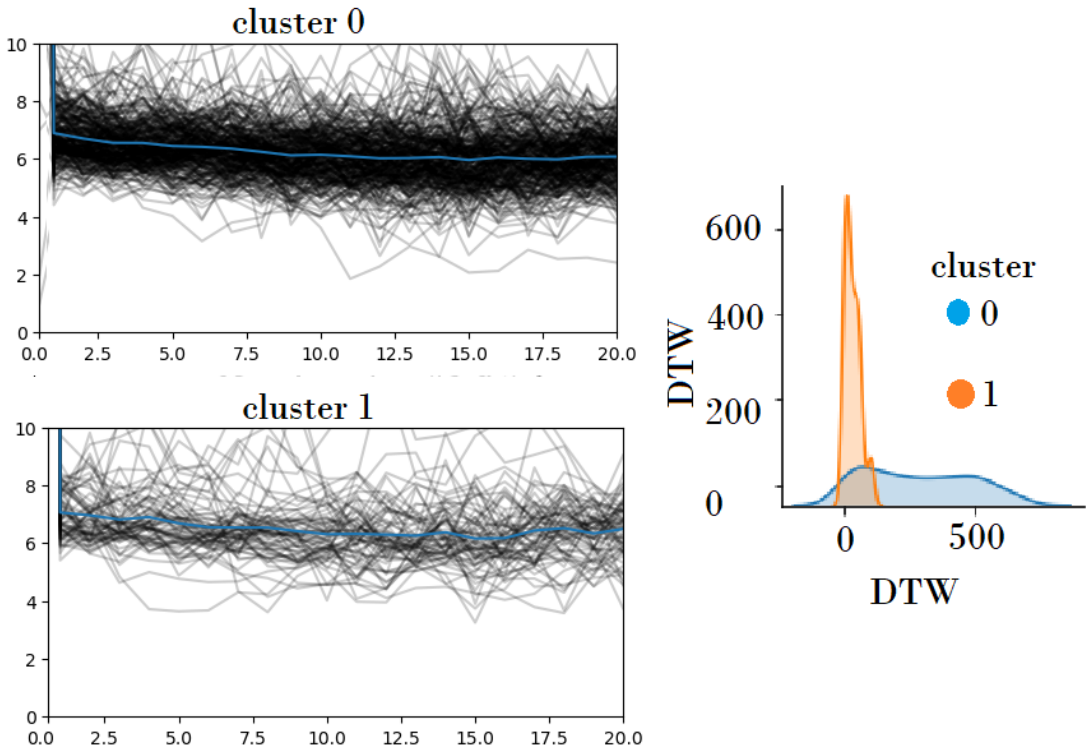

Fig. 3.11 distribution des mesures *DTW*, entre les séries et leur représentant, par cluster de series temporelles d'oxygène dissous

affiche aussi la valeur d'entropie de ces distances.

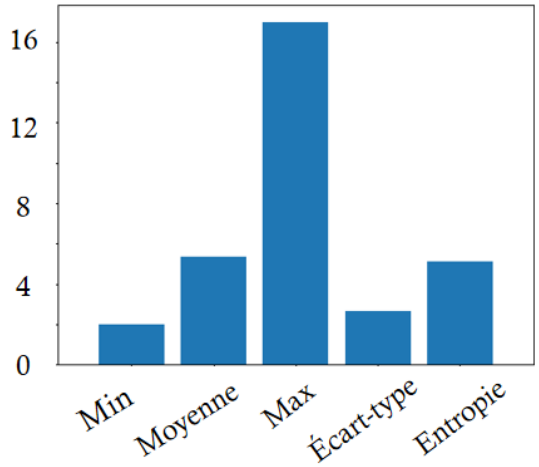

Fig. 3.12 Données statistiques des mesures de distances *DTW*, entre les séries et leurs représentant, pour le cluster 1 de température

La figure 3.13 présente pour le cluster 0 de l'OD, les mêmes statistiques (moyenne, écart type, minimal, maximal, entropie) des distances. Pour l'OD, l'entropie des distances *DTW* intra-cluster, est inférieure à leur écart-type, contrairement aux statistiques des distances *DTW* des clusters de température.

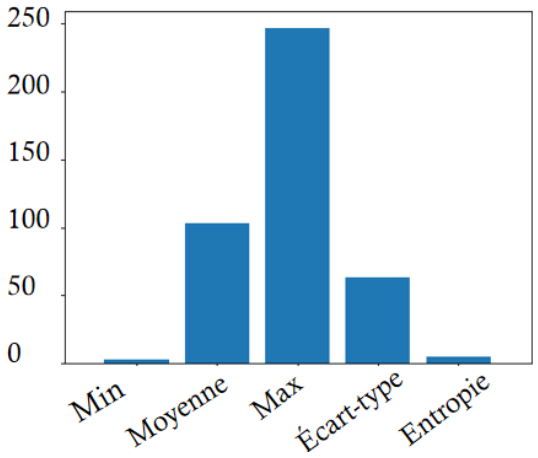

Fig. 3.13 Données statistiques des mesures de distances *DTW*, entre les séries et leurs représentant, du cluster 1 pour l'oxygène dissous

Ainsi lorsque l'écart-type est inférieur à l'entropie, intra-cluster, l'évolution des variations des séries temporelles est moins importante, les séries ont des tendances homogènes. Il y a par exemple une tendance décroissante pour le cluster 0 de température et croissante pour le cluster 1. Et pour ces clusters, les statistiques montrent un écart type des distances, inférieur à leur entropie. Cela ne se vérifie pas pour les séries d'OD dont l'évolution des formes n'est visiblement pas homogène par cluster.

On remarque enfin, entre les clusters de ces deux variables environnementales, que les distributions des distances *DTW*, (figure 3.11 et 3.10) sont plus symétriques, lorsque l'écart-type est inférieur à l'entropie (ce qui est le cas pour les clusters de température contrairement aux clusters d'OD). La nouvelle mesure de dispersion proposée s'inspire de cette différence.

3.5 Présentation de la nouvelle approche pour le clustering de séries temporelles monovariées

La stratégie de la méthode mono-variée proposée, permet de définir automatiquement un nombre k optimal de clusters selon un critère de dispersion des individus autour du représentant, du cluster. Notre méthode utilisera des distances prenant en compte le décalage temporelle, et nous définirons une nouvelle mesure de dispersion qui va permettre de regrouper les séries d'amplitudes similaires ou proches selon cette mesure de dispersion. Contrairement à la plupart des méthodes qui normalisent les données, notre approche s'applique tant sur les séries temporelles normalisées que les séries temporelles brutes.

3.5.1 Clustering de séries temporelles monovariées : X-MeansTS

En entrée de notre algorithme, un paramètre supplémentaire fixant le nombre maximum d'itérations sera fourni en cas de non convergence de la méthode (comme pour la méthode de k-means). Notre approche se concentre plutôt sur une nouvelle stratégie d'affinage robuste des clusters. Elle entraîne la génération automatique du nombre de k de clusters en fonction d'une nouvelle mesure de dispersion calculée à partir de la mesure de distance utilisée. Nous testerons notre approche avec différentes mesures de distance (par ex. les mesures dérivées de *DTW*).

Le principe de la méthode est d'affiner chaque cluster en revisitant chaque instance à partir d'un nombre minimal, de clusters, fixé initialement à nb_min_clust et d'un ensemble de critères, que nous définirons ensuite. Les instances ne vérifiant pas les critères, par rapport à leur classe d'appartenance, sont mis dans une classe "rejet". La figure 3.14 présente schématiquement le principe de la méthode. On itère le principe sur cette classe rejet (considérée comme nouvel ensemble de séries à clusteriser) jusqu'à ce qu'un ensemble de conditions d'arrêt soit vérifié.

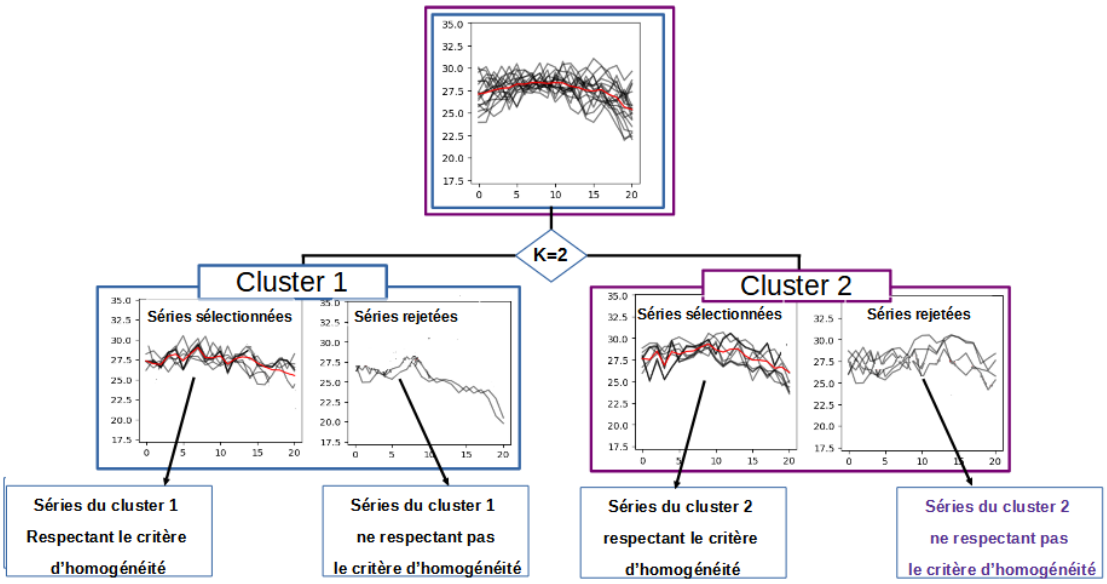

Fig. 3.14 Stratégie de la nouvelle approche

Nous détaillerons dans les sections suivantes les critères de sélection des instances et les conditions d'arrêt. La section suivante apportera une démonstration mathématique de l'amélioration des performances en terme d'homogénéité intra-cluster, grâce à critère.

3.5.2 La nouvelle mesure de dispersion

On note $Dist(C_i) = \{d_{i_1}, \ldots, d_{i_{m_i}}\}$ l'ensemble des valeurs prises par la mesure $Dist$ entre les instances de C_i et leur représentant R_{C_i}. Soit $\sigma(C_i)$ l'écart-type calculé sur la

distribution des valeurs prises par $Dist(C_i)$ et $E(C_i)$ sa mesure d'entropie :

$$\sigma(C_i) = \sqrt{\sum_{k=1}^{m_i}(d_{i_k} - \overline{d_i})^2} \tag{3.3}$$

$$E(C_i) = -\sum_{k=1}^{m_i} P(d_{i_k}) \times log(P(d_{i_k})). \tag{3.4}$$

où $\overline{d_i}$ est la moyenne de $Dist(C_i)$.
On définit alors la mesure de dispersion, notée $disp(C_i)$ par :

$$disp(C_i) = \frac{\sigma(C_i)}{E(C_i)} \tag{3.5}$$

Propriété de la nouvelle mesure de dispersion $disp(C_i)$: Nous démontrerons ici que la nouvelle mesure de dispersion est liée à la mesure de l'homogénéité.

Corollaire 3.5.1. *Si $\sigma(C_i)$ tend vers 0 alors $h(C_i)$ tend vers sa valeur maximale 1.*

On rappelle, pour les besoins de la démonstration, le calcul de l'homogénéité d'un cluster : Soit *L* un ensemble de labels réels $L = \{Li|i = 1, \ldots, m\}$, on désigne par a_{ij} le nombre d'instances de label i affectées au cluster j. L'homogénéité h calculée pour un ensemble de clusters C est définie comme :

$$h(C) = \begin{cases} 1 & \text{If } H(L,C) = 0 \\ 1 - \frac{H(L|C)}{H(L)} & \text{else.} \end{cases}$$

où $H(L|C) = -\sum_{c=1}^{|C|}\sum_{l=1}^{|L|}\frac{a_{lc}}{N}\log\frac{a_{lc}}{\sum_{l=1}^{|L|}a_{lc}}$ and $H(L) = -\sum_{l=1}^{|L|}\frac{\sum_{c=1}^{|C|}a_{lc}}{m}log\frac{\sum_{c=1}^{|C|}a_{lc}}{m}$.
Démonstration : La mesure de dispersion $disp(C_i)$ traduit la variabilité à l'intérieur d'un cluster. Plus $disp$ est petite, plus la variabilité, des individus autour du représentant est faible.
Plus cette variabilité intra-cluster est faible, plus la probabilité d'avoir un groupe d'individus homogène est élevé. Une variabilité nulle peut donc se traduire par un cluster contenant des individus du même label. Supposons que les individus d'un clusters C_i soient associés à un label unique. Ainsi la courbe de distribution de $Dist(C_i)$ pourrait être d'autant plus "resserrée" autour de son axe de symétrie (du représentant) que l'écart type serait faible et que l'entropie serait importante. En effet l'entropie par définition, augmente avec l'équilibre des probabilités des valeurs dans une distribution. De plus les valeurs de la distribution de $dist(C_i)$ seraient identiques si l'écart serait nulle. On sait que : $\max(h(C)) \Leftrightarrow \min(\sum_{i=1}^{|C|} disp(C_i))$,
$\max(h(C)) \Leftrightarrow \min(\frac{H(L|C)}{H(L)})$ et $\min(\sum_{i=1}^{|C|} disp(C_i)) \Leftrightarrow \min(\sum_{i=1}^{|C|}\frac{\sigma(C_i)}{E(C_i)})$

Le domaine de définition : $\forall C_i \in C$, $\sigma(C_i) \in \mathbb{R}^+$, $E(C_i) \in \mathbb{R}^+$ et $\frac{\sigma(C_i)}{E(C_i)} \in \mathbb{R}^+ \Leftrightarrow \min(H(L_j|C_i)) = \min(\sigma(C_i))$.
Or $\min(disp(C_i)) \Leftrightarrow \min(\frac{\sqrt{\sum_{k=1}^{m_i}(d_{i_k}-\overline{d_i})^2}}{-\sum_{k=1}^{m_i} P(d_{i_k})\times log(P(d_{i_k}))})$ avec comme contrainte suivante $E(C_i) > 0$.
Pour chaque cluster C, minimiser $disp(C)$ revient à minimiser $\sigma(C)$ et à maximiser $E(C)$.

Si $\sigma(C) = 0$ alors $\forall k \in \{1, \ldots, m\}, d_{kC} = \overline{d_C}$ ($\overline{d_C}$ = moyenne des distances des instances à leur représentant du cluster C). Cela signifie qu'il y a au plus une instance par cluster (ou qu'il y a plusieurs instances identiques).

Soit a_{lC} le nombre d'instances avec le label l dans le cluster C; mais $E(C) > 0$ alors $\log(P(d_{kC})) \neq 0$. Cela signifie que $a_{lC} > 1$ et $|L| > 1$.

$\min H(L|C) = \min(-\sum_{c=1}^{|C|}\sum_{l=1}^{|L|}\frac{a_{lc}}{N}\log\frac{a_{lc}}{\sum_{j=1}^{|L|}a_{jc}})$ avec $d_{iC} = \overline{d_C}$, on a $a_{lC} \in \{1, |L|\}$ et comme $|L| > 1$ alors $a_{lc} = |L|$ et $-\sum_{c=1}^{|C|}\sum_{l=1}^{|L|}\frac{|L|}{N}\log\frac{|L|}{\sum_{l=1}^{|L|}1} = 0$ avec $H(L|C)$ = 0 par conséquent $h(C)$ = 1.

■

3.5.2.0.1 Théorème de Tokotoko-Govan La mesure de dispersion $disp(C))$ permet d'améliorer l'homogénéité des clusters.

Le paragraphe suivant présentera la stratégie de la méthode d'affinement des clusters en sortie de méthodes de clustering existantes.

3.5.3 Principe de l'algorithme *X-MeansTS*

Notre approche X-MeansTS nécessite la donnée de 3 paramètres liés à des seuils à fixer en entrée de l'algorithme :

1. nb_min_clust : le nombre de clusters initiaux, générés par une approche de clustering existante (*KmeansTS*).
2. nb_min_inst : le nombre minimum d'instances admis par cluster
3. s_d : valeur maximale de la mesure de dispersion $disp$. En effet, elle traduit la variabilité intra-groupe.

En sortie de notre algorithme nous obtenons un nombre de clusters déterminé automatiquement (selon les critères appliqués) et une classe rejet, notée CR, contenant les

instances ne vérifiant pas le critère de dispersion.
De manière générale, l'approche s'exécute selon le schéma de la figure 3.15 :

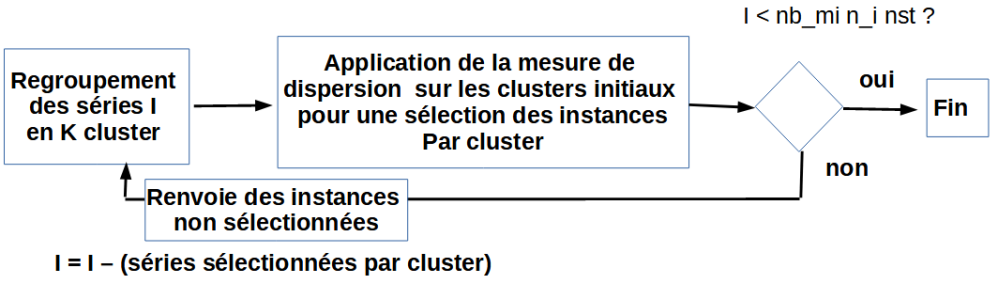

Fig. 3.15 Concept général de la méthode de clustering mono-varié *XmeansTS*

L'idée générale de notre méthode est de partir d'un groupement initial avec un nombre minimum de cluster nb_min_clust. Le groupement initial est construit par une méthode de clustering existante (par ex. *K-MeansDTW*, *K-shape*, etc.). Le critère d'homogénéité, en l'occurrence notre mesure de dispersion, est appliqué à chaque cluster. Les instances de ce cluster sont analysées une par une. Celles qui ne vérifient pas le critère d'homogénéité sont affectées à une classe rejet, sinon elles restent dans le cluster. On itère le processus complet sur la classe rejet, tant que cette classe n'est pas vide (ou contient un nombre minimum d'instances). Durant le processus, les clusters générés initialement, ayant un nombre d'instances inférieur au seuil fixé nb_min_inst sont supprimés et leurs instances sont affectées à la classe rejet. De manière plus détaillée,

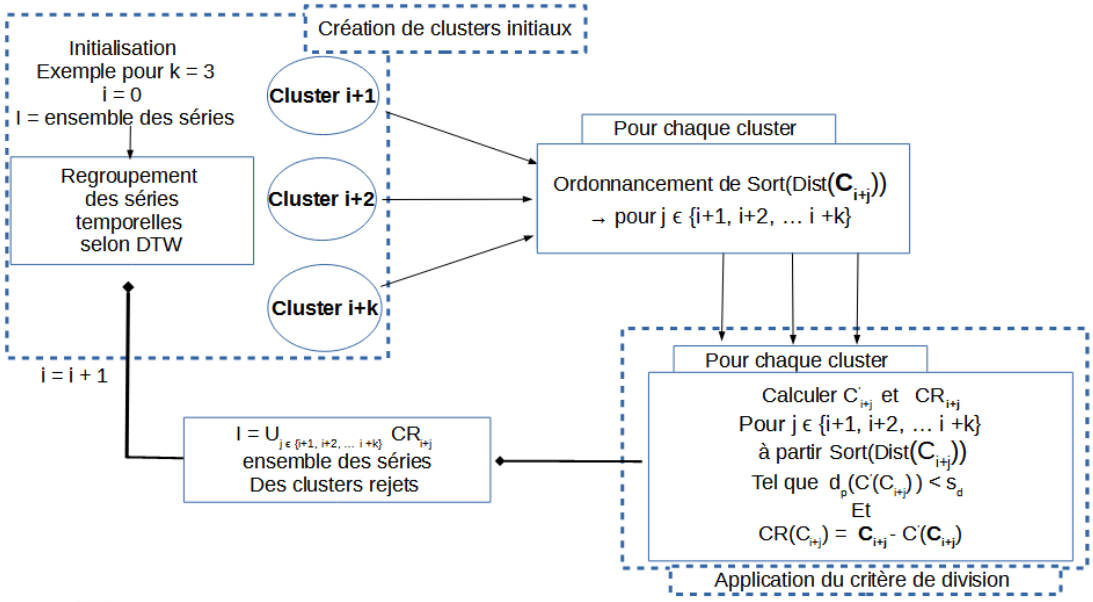

Fig. 3.16 Principe de la méthode de clustering mono-variée *X-meansTS*

l'algorithme se déroule selon la figure 3.16 dont les étapes sont présentées ci-dessous :

1. **Étape initiale (définition des clusters initiaux) :** Les instances d'un ensemble de séries temporelles monovariées I, sont partitionnées en un nombre nb_min_clust minimal de clusters. Pour créer ces clusters, on applique par exemple l'algorithme *K-MeansTS* ([70] avec $k = nb\_min\_inst$ et $Dist$ la mesure de distance utilisée (qui peut être DTW, etc.).

2. **Étape d'affinement des clusters par l'application du critère de dispersion :** Dans cette phase 2, on applique la mesure de dispersion pour chaque cluster C_i. Si le nombre d'instances du cluster initial C_i est inférieur à nb_min_inst, alors ce cluster est supprimé et ses instances sont affectées à la classe rejet CR. Sinon, on sélectionne les instances (séries temporelles) que l'on affecte à un nouveau cluster C'_i, une par une, dans l'ordre croissant de leur distance au représentant du cluster, tant que sa dispersion $disp(C'_i)$ (qui est mise à jour au fur et à mesure) reste inférieure au seuil fixé s_d. Les instances qui ne vérifient pas ce critère sont affectées à la classe rejet CR.

3. **Étape d'application du critère d'arrêt** : Si le nombre d'instances de la classe rejet est supérieur à nb_min_inst, alors on répète le processus à partir de l'étape initiale en prenant comme nouvel ensemble à clusteriser CR la classe rejet. L'algorithme s'arrête si le cluster rejet CR est vide, ou que le nombre d'instances à traiter CR est inférieur strictement à nb_min_inst.

La première étape est basée sur un algorithme de clustering des séries temporelles existant avec une mesure de distance adaptée aux séries temporelles.

Le calcul de la mesure de dispersion d'un cluster C_i nécessite au moins deux valeurs dans $dist(C_i)$. Le paramètre nb_min_inst qui est le nombre minimum d'instances exigé dans chaque cluster nous permet de fixer les premières instances initiales dans le nouveau cluster C'_i. Afin de déterminer ces instances, les $Dist(C_i)$ sont ordonnées et sauvegardées dans $Sort(Dist(C_i)) = \{v_1, v_2, \ldots, v_m\}$ avec $\forall\ i < j$, $v_i \leq v_j$ (voir algorithme 8). On intègre dans C'_i les premières nb_min_inst instances de la liste triée $Sort(Dist(C_i))$. Si $disp(C'_i) \leq s_d$ alors les autres instances sont ajoutées une à une dans C'_i, tant que le critère reste vrai. Sinon les instances qui ne vérifient pas le critère sont donc mises dans le cluster de rejet. La valeur $disp(C'_i)$ est mise à jour à chaque fois qu'une instance est ajoutée.

L'algorithme 9 présente ainsi le déroulement de la méthode basée sur les critères décrits précédemment ($s_d, nb\_min\_inst, nb\_min\_clust$). À chaque appel de l'algorithme, un nouvel ensemble de nb_min_clust clusters est généré et seront affinés en utilisant la mesure de dispersion, pour la sélection de séries respectant le seuil de dispersion paramétré. Les individus non sélectionnés, par cluster selon ce seuil (paramétré) sont intégrées dans le cluster rejet. La méthode est appliquée à l'ensemble des séries su

Algorithm 8 ApplyCriteria

Input:
- $C = \{s_1, s_2, \ldots, s_m\}$
- $Dist$: la mesure de distance choisie
- s_d seuil de dispersion
- nb_min_inst le nombre minimum d'instances

Output:
- CR : la classe rejet
- C' : le cluster modifié

1: Calcul de $Sort(Dist(C)) = \{v_{i_1}, v_{i_2}, \ldots, v_{i_m}\}$ avec $\forall j < r$, $v_{i_j} \leq v_{i_r}$
2: $C' = \{s_{i_1}, s_{i_2}, \ldots, s_{i_k}\}$ où $k = nb\_min\_inst$
3: h = 0
4: **while** $d_p(C') < s_d$ and $i < m$ **do**
5: $\quad C' = C' \cup \{s_{k+h}\}$
6: $\quad$ h = h+1
7: **end while**
8: $CR = C - C'$
9: **return** $\{C', CR\}$

cluster rejet à l'étape suivante, tant que le nombre de séries le composanest supérieur à nb_min_inst. La procédure $[C, Dist(C)] = CreateInitialsClusters(T, nb\_min\_clust)$ de l'algorithme 9 fait appel à une méthode de clustering existante telle que la méthode (*K-Shape*, *KmeansDTW*...).

Remarquons qu'au premier appel ou au cours des appels récursifs de l'algorithme 9, il peut être assigné au cluster rejet CR les mêmes instances indéfiniment selon les valeurs des paramètres d'entrées. En effet, si, par exemple, le seuil de dispersion s_d est très faible alors des instances affecteront la valeur de la mesure de dispersion qui devient supérieur au seuil. Ces instances n'intégreront aucun cluster. Nous introduisons alors, dans l'algorithme 9, un paramètre supplémentaire $recursifCpt$ un compteur de récursivité et qui jouera le rôle d'un autre critère d'arrêt dés qu'il atteint le nombre d'itérations choisi (paramétrable par l'utilisateur).

Il est ainsi possible que le nombre de clusters soit inférieur à nb_min_clust voire qu'il n'y ait pas de cluster. Cette possibilité intervient lorsque la méthode $ApplyCriteria$ ne trouve aucune instance vérifiant le critère de dispersion, dans chacun des clusters initiaux. Cet état est lié à une faible valeur du seuil de dispersion. Néanmoins, l'augmentation du seuil intégrera des instances qui sont éloignées du représentant et entraînera la création d'un cluster avec une variabilité importante.

L'algorithme ainsi fournit comme résultat, un ensemble de clusters dont le nombre est déterminé automatiquement et un cluster rejet contenant des instances qui seront qualifiées d'instances isolées.

Algorithm 9 X-MeansTS

Input:
- $T = \{s_1, s_2, \ldots, s_p\}$ un jeu de données temporelles
- nb_min_clust : nombre minimum de cluster à l'étape initiale
- s_d : seuil de dispersion
- $nbMaxIter$: nombre maximum d'itérations
- $nbClust$: nombre de clusters obtenus initialisé à 0
- $recursifCpt$: nombre d'appels récursifs
- nb_min_inst nombre minimum d'instances

Output: - C_f ensemble de clusters

if $nb\_min\_clust > p$ **then**
  $\{C, Dist(C)\} = CreateInitialsClusters(T, nb\_min\_inst)$
  for i in 1 to nb_min_clust **do**
    $\{C'$, $CR\}$ = ApplyCriteria(C_i, $Dist$, s_d, nb_min_inst)
    if $C' \neq \emptyset$ **then**
      $C_f[nbClust] = C'$
      $T = T - C'$
      $nbClust = nbClust + 1$
    else
      $recursifCPT = recursifCpt + 1$
    end if
  end for
  if $recursifCpt < nbMaxIter$ **then**
    $XmeansTS(T, nb\_min\_clust, s_d, nbClust, nbMaxIter, recursifCpt, nb\_min\_inst)$
  end if
end if
return C_f

3.6 Validation de l'algorithme X-meansTS

La méthode de clustering de séries temporelles monovariées *X-MeansTS* a été testée sur différents jeux de données Benchmark *UEA and UCR* disponible en ligne [7, 21] et sa performance a été comparée à la performance de la méthode *K-shape* et de la méthode *K-meanTS*, les plus connus. La section suivante (section 3.6.1) présentera les résultats de ces tests afin de valider l'efficacité de cette nouvelle approche en l'occurrence pour des données complexes ayant un nombre de clusters très élevé. Le chapitre 7 fournira des résultats qualitatives et des interprétations de ces résultats sur le jeu de données réelles de la filière d'aquaculture Calédonienne.

3.6.1 Expérimentations de l'algorithme *X-MeansTS*

Les jeux de données de tests, des archives *UEA and UCR* [7, 21] contient actuellement près de 200 jeux de données faisant références à des domaines variés (sport, science, littérature, industrie...).

Pour chaque jeu de données que nous avons utilisé, les informations suivantes sont fournies :

- le nom du jeu de données
- le nombre de séries temporelles
- le label de chaque série
- la longueur des séries

Le label des séries est évidemment utile dans le cadre d'une classification supervisée pour calculer par exemple, la précision du modèle. Dans le cas de notre méthode de clustering, il servira à calculer sa performance, en fonction de l'homogénéité des clusters.

Afin d'avoir des performances établies sur un ensemble de jeux de données très hétérogènes, les tests ont été réalisés sur 20 jeux de données avec des séries de longueurs variées et un nombre de classes différent (cf. table 3.1). Pour tester notre nouvelle méthode *X-meansTS* sur des données complexes, Les 7 derniers jeux de données du tableau 3.1 ont un nombre plus élevé de classes (entre 24 et 60). De plus, certains d'entre eux contiennent des séries de plus grande taille (1250 points) et/ou un nombre assez élevé de séries (par exemple 7800 séries pour les données Crop). Une comparaison des résultats obtenus sur ces 7 jeux de données par rapport aux résultats des méthodes (*K-MeansTS, K-shape*), est présentée.

Recherche automatique des seuils des paramètres en entrée de la méthode *X-meansTS*: Pour chaque jeu de données considéré, l'algorithme a été expérimenté en faisant varier principalement le seuil s_d de la mesure de dispersion $disp$, et le nombre nb_min_clust de clusters initiaux. Concernant le paramètre nb_min_inst du nombre minimum d'instances par cluster, il a été fixé selon le jeu de données. Pour assurer l'obtention de tous les clusters, le paramètre nb_min_inst, quand à lui, sera fixé à un nombre inférieur ou égal au nombre d'individus de la classe qui en possède le moins.

Des combinaisons de couples de valeurs pour les paramètres s_d et nb_min_clust, ont été déterminées, pour chaque jeu de données, à partir d'une liste de valeurs pour chaque paramètre. Ces listes ode valeurs ont été générées selon la même approche pour tous les jeux et de la manière suivante : la liste des seuils est obtenue à partir des résultats de différents clustering obtenus par une méthode existante telle que *K-meansDTW* par exemple. Étant donné que le nombre de clusters souhaité est configurable avec cette méthode, la méthode a été paramétrée de manière à obtenir des résultats pour un nombre de clusters variant entre, au minimum 2, et, au maximum, le double du nombre de classes réelles du jeu.

Par exemple, pour le jeu de données "Car", le nombre maximal est fixé à 4 puisque le nombre de classes est de 2. Ainsi pour ce jeu, l'algorithme *KmeansTS* est exécuté 3

fois, c'est à dire, en faisant varier le nombre *k* de clusters de 2 à 4.
On obtient ainsi $2 + 3 + 4 = 9$ clusters pour ce jeu de données. Ces clusters générés permettent d'avoir, par jeu de données, un intervalle de valeurs de seuils de dispersion. En effet, en fonction de la variabilité des séries dans les différents jeux, les seuils de la mesure de dispersion seront différents (La mesure de dispersion permettant d'obtenir des clusters avec des séries évoluant sur des amplitudes de plus en plus proches, lorsque la mesure se rapproche de 0).

Dès lors que, par jeu de données, la liste de seuils (de dispersion) est déterminée, on génère la liste de valeurs du paramètre nb_min_clust, c'est à dire du nombre de clusters initiaux. Pour cela, par jeu, une liste de valeurs est générée, allant de 2 jusqu'au nombre de classes réel. L'algorithme *X-MeansTS* est ainsi exécuté avec chaque valeur de la liste de seuils, et pour chaque exécution, on fait varier le paramètre nb_min_clust selon sa liste de valeurs.

De plus, cette configuration d'exécutions de l'algorithme *X-MeansTS* par jeu de données, est répétée pour différentes mesures de distances, entre deux séries temporelles, utilisables avec l'algorithme. Ces mesures sont principalement les mesures dérivées de DTW lorsque *K-MeansDTW* est utilisée dans *X-MeansTS* pour générer les clusters initiaux.

Pour expliquer le protocole expérimental pour chaque jeu de données avec notre stratégie de recherche automatique des seuils, prenons l'exemple du jeu de données *'Car'*. Pour ce jeu de données, on obtient par exemple une liste de 9 valeurs de seuils différents, calculés à partir des 9 clusters pré-cités. À partir de cette liste de seuils, on exécute *X-MeansTS* avec chacun des éléments de la liste, comme paramètre d'entrée s_d. Ainsi, pour chaque seuil, on itère sur le paramètre nb_min_clust i.e le nombre de clusters initiaux en le faisant varier de 2 à 4 clusters. On obtient $9 * (2 + 3 + 4) = 81$ clusterings qui fournissent chacun un nombre optimal de clusters (voire aucun cluster) selon les couples de valeurs de s_d et nb_min_clust.

Chaque fois que *X-MeansTS* fourni un résultat pour chaque jeu de données testé, pour ces combinaisons de paramètres, *K-MeansDTW* et *K-Shape* ont été exécutés (plusieurs fois) avec, en entrée, le nombre de clusters k fourni en sortie par *X-MeansTS*, afin de pouvoir comparer leurs performances à celle de *XmeansTS* (section 3.6.2). Cette façon d'obtenir des résultats avec ces 3 méthodes de clustering (*X-MeansTS*, *K-MeansDTW* et *K-Shape*) a été appliquée, comme énoncé, en utilisant différentes mesures de distance (*DTW + multiscale, DTW + itakura, FastDTW...*). Les résultats de la méthode *X-MeansTS* ont été évalués et comparés à ceux des méthodes *K-MeansDTW* en utilisant les mêmes mesures de distance. *X-MeansTS* obtient d'excellents résultats sur ces jeux de données (selon la mesure de l'indice Rand [157]). Au total, plusieurs milliers de tests ont été générés, et leurs performances ont été décrites en fonction de la mesure de distance utilisée.

Dataset	nb instance	nb classe	taille série
BirdChicken	20	2	512
Car	60	4	577
Computers	250	2	720
DiatomSizeReduction	16	4	345
FaceFour	24	4	350
Fish	175	7	463
Ham	109	2	431
Herring	64	2	512
LargeKitchenAppliances	375	3	720
Lightning7	70	7	319
Meat	60	3	448
OliveOil	30	4	570
OSULeaf	200	6	427
RefrigerationDevices	375	3	720
ScreenType	375	3	720
SmallKitchenAppliances	375	3	720
Earthquakes	322	2	512
ToeSegmentation2	36	2	343
Yoga	300	2	426
Adiac	390	37	176
Crop	7200	24	46
EOGVerticalSignal	362	12	1250
FiftyWords	450	50	270
GestureMidAirD1	208	26	360
NonInvasiveFetalECGThorax1	1800	42	750
ShapesAll	600	60	512

Table 3.1 Description des jeux de données

Les principales métriques de performance utilisées pour évaluer les clustering sont **ARI (Adujsted random index)** [157] et la **V-Mesure** ([145]). La V-Mesure a été choisie car, elle a été définie pour répondre aux deux critères (les plus utilisés pour évaluer la qualité de la méthode de clustering) suivants : l'homogénéité, et la complétude des clusters.

Les tests ont été effectués avec un processeur Intel i7-6700HQ, 2.60GHz, le temps de calcul est comparable à celui de *K-Shape* et *K-MeansDTW* sur l'ensemble des jeux de données utilisées.

3.6.2 Résultats

Pour la recherche des clusters initiaux nécessaire à l'algorithme *X-MeansTS*, nous utiliserons la méthode *K-Shape* et la méthode *K-MeansDTW* avec différentes mesures de distance dérivées de *DTW*.

Les définitions qui suivent précisent comment les mesures de performances ($V\_measure$ et ARI) ont été utilisées :

- **ARI_KMeansDiff** (resp. **ARI_ShapeDiff**) mesure la différence de l'ARI entre *X-MeansTS* et *K-MeansTS* (resp. **K-Shape**). Si $ARI\_KMeansDiff > 0$ alors les clusters fournis par *X-MeansTS* se rapprochent d'autant plus du jeu de données réel, que ceux fournis par *K-MeansTS*.

- **Vmeas_KMeansDiff** (reps. **Vmeas_K-ShapeDiff**) mesure la différence de $V\text{-}mesure$ entre les méthodes *X-MeansTS* (resp. *K-Shape*) et *K-meansTS*.
  Si $Vmeas\_KMeansDiff > 0$ (resp. $Vmeas\_ShapeDiff > 0$), alors les clusters fournis par *X-MeansTS* sont plus homogènes que la méthode *K-MeansTS*.

Scénarios de comparaison : La comparaison des performances par ces mesures a été réalisée à la fois sur des jeux de données contenant moins de 10 classes, et un nombre de séries inférieurs à 400, et sur des jeux avec au nombre de classe allant de 10 à 60 et un nombre de séries allant de 360 à 7200. Dans le premier cas, nous nommerons par la suite les jeux avec moins de 10 classes, des 'jeux simples', dans le deuxième cas des 'jeux particuliers'. Pour cette comparaison, plusieurs milliers de tests ont été réalisés en modifiant les paramètres d'entrée de la méthode *X-meansTS* d'après les scénarios de combinaisons de paramètres présentés dans la section ..... A chaque combinaison de paramètre, différentes mesures de distances dérivées de *DTW* (*fastDTW*, DTW+*itakura*....) sont aussi testés. A chaque test, le nombre de clusters obtenus, sert de paramètre *K* en entrée de la méthode *K-meansDTW*, et la mesure de distance utilisée lors du test par *X-meansTS*, est intégrée (dans *K-meansDTW*). On obtient ainsi, d'après les scénarios (de combinaisons de paramètres) plus de 1500 tests au total par *X-meansTS* sur l'ensemble des 26 jeux de données du tableau 3.1. Et par conséquent plus 1500 tests également effectués par la méthode *K-meansDTW* et la méthode *K-Shape*. Dans l'ordre de présentation des résultats, nous étudierons ces mesures de comparaison dans les cas suivants :

- lorsque qu'un même nombre de cluster générés, par les 3 approches, est différent du nombre réel de classe. Cela permet d'étudier, pour un même nombre de cluster généré par les méthodes, la différence de qualité en terme d'homogénéité, entre ces approches lorsque que leurs nombres de clusters sont soient inférieurs, soient supérieurs au nombre réel de classe par jeu de données.

- en fonction des mesures de distances dérivées de *DTW* utilisées (*fastDTW*, DTW+*itakura*, ...), en séparant les jeux (de données) simples des jeux particuliers.

Comparatif général de la qualité des clusters entre *X-meansTS*, *K-meansDTW* et *K-Shape* : La figure 3.17 affiche les métriques qui calculent la différence de qualité

entre les résultats de *Xmeans-TS* et *Kmeans-DTW* sur la base d'un même nombre de cluster i.e *ARI_KMeansDiff* et *Vmeas_KMeansDiff*). La figure 3.17 affiche les résultats du clustering pour lesquels les nombres de clusters générés par ces deux méthodes, sont au plus, inférieur de 3 ou supérieur de 4, du nombre réel de classe. Cette différence de nombre, est noté *kdiff* dans la figure. D'après cette figure 3.17 les clusters obtenus par *X-MeansTS* sont souvent plus homogènes et plus similaires (cf. ARI) que *K-MeansDTW*. On note en effet un nombre supérieur de valeurs positives pour *ARI_KMeansDiff* et *Vmeas_KMeansDiff*. Et l'étendue des valeurs positives de ces deux métriques est plus important, que leurs valeurs négatives. La différence de qualité des clustering est donc plus importante lorsque c'est notre méthode *Xmeans-TS* qui, sur la base d'un même nombre de clusters, génère des clusters plus homogènes que la méthode *Kmeans-TS*. De la même manière, la figure 3.18 présente une comparaison des résultats, via la différence des mesures de performances *V-mesure* et *ARI*, entre notre méthode *X-MeansTS* et la méthode *K-Shape*. Contrairement à la comparaison précédente de la qualité des clusters entre *Xmeans-TS* et *Kmeans-TS*, la différence entre *Xmeans-TS* et *K-Shape*, montre des résultats en faveur de *K-shape*. On note, en effet, un nombre supérieur de valeurs négatives d' *ARI_KMeansDiff* et de *Vmeas_KMeansDiff*.

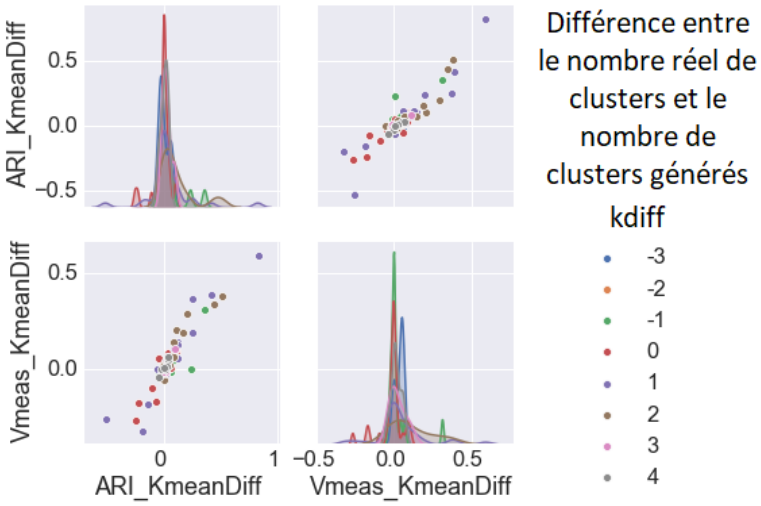

Fig. 3.17 *V-mesure* et *ARI* en fonction de la différence entre le nombre de classes réel et le nombre de clusters obtenus

Afin de déterminer si la mesure de distance choisie, impacte fortement les valeurs de ARI et $Vmesure$, nous avons analysé la différence de qualité des clusters en fonction des mesures des distances *DTW* utilisées et notemment par *X-MeansTS* et *K-MeansDTW*. En plus de cette comparaison en fonction de ces mesures, les résultats seront présentés en fonction des jeux simples et des jeux particuliers.

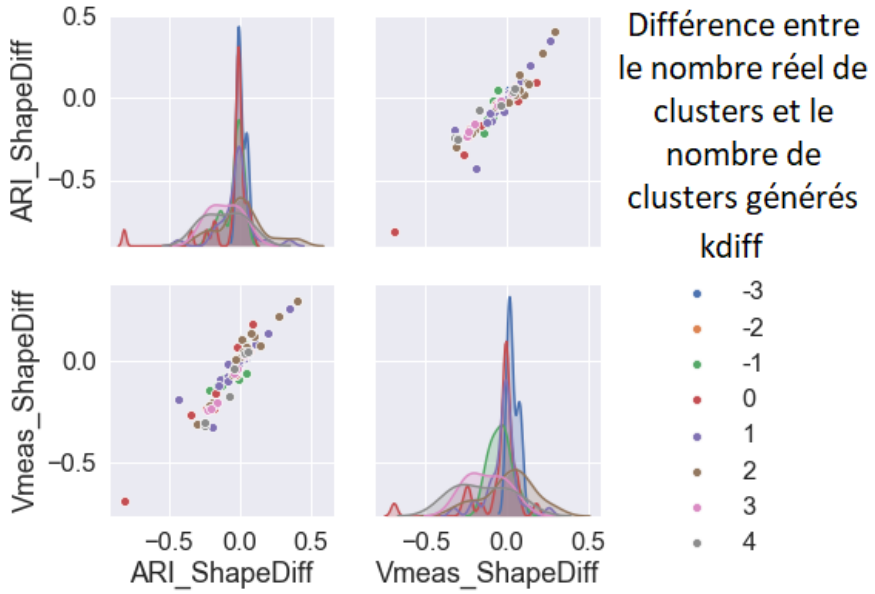

Fig. 3.18 *V-mesure* et *ARI* en fonction de la différence entre le nombre de classes réel et le nombre de clusters obtenus.

Pour cela, les graphiques qui suivront, afficheront des valeurs normalisées de ces deux mesures (de différence de qualité) proposées (*Vmeas_Kmeansdiff*, *ARI_Kmeansdiff*...). Par exemple si la valeur de V-mesure de notre méthode *X-meansTS* vaut 0.4 et celle de *K-meansDTW* vaut 0.1 alors le taux sera positive et vaudra 0.25. Cela signifiera que l'homogénéité de *X-meansTS* est 25% plus élevée que celle de méthode *K-meansDTW* sur la base d'un même nombre de clusters générés par ces deux méthodes. Réciproquement, si la valeur de *V-mesure* de notre méthode *X-meansTS* vaut 0.1 et celle de *K-meansDTW* vaut 0.4 alors le taux sera négative et vaudra 0.25. Cela signifiera que l'homogénéité de *X-meansTS* est 25% moins élevée. Sur les milliers de tests énoncés le calcul de ce pourcentage est fait et on affichera ainsi des boxplots de ces pourcentages. Cela permettra d'afficher des taux, obtenus sur ces milliers de tests, en terme de différence de qualité entre notre méthode *X-meansTS* et les méthodes *K-meansDTW* et *K-Shape*.

Comparatif de la qualité des clusters entre *X-meansTS* vs *K-meansDTW* en fonction des jeux de données simples : Sur les milliers de tests énoncés précédemment, la figure 3.19 compare la différence de qualité entre les clusters générés par *X-meansTS*, et la méthode *K-meansDTW* sur les jeux simples. La figure montre cette différence de qualité, en fonction des mesures dérivées de *DTW* en abscisse, utilisées par les deux méthodes pour les tests. Dans cette figure la figure 3.19 deux graphiques sont affichés. Le graphique de gauche affiche pour différentes mesures de distances, le taux de *Vmeas_KMeansDiff*, et celui de droite le taux *ARI_KMeansDiff*. Sur la base d'un

même nombre de cluster (générés par *X-meansTS* et *K-meansDTW*), les différences de qualité déterminés par *Vmeas_KMeansDiff*, et *ARI_KMeansDiff* ont donc été calculés. Comme énoncé, ces valeurs ont été normalisées pour obtenir, une distribution du taux de différence de qualité entre notre méthode et *K-meansDTW*. D'après cette figure 3.19, *X-MeansTS*, selon les taux de *Vmeas_KMeansDiff*, est en moyenne 2% plus homogènes que *K-MeansDTW* sur les jeux de données simples (et pouvant atteindre des différences de taux qualité d'environ 10% en faveur de *X-MeansTS*). Les mesures de distance *DTW* combinées avec le parallélogramme de *Sakoechiba* et la bande d'*Itakura* ont des distributions beaucoup plus étalées concernant les valeurs liées aux différences d'homogénéité et de similarité entre *X-meansTS* et *K-meansDTW*) (i.e *Vmeas_KMeansDiff* et *ARI_KMeansDiff*). En général, les clusters obtenus par *X-MeansTS* sont souvent plus homogènes et plus similaires aux vrais clusters des jeux de données 'simples' (cf. *ARI*) que *K-MeansDTW*. Avec l'approche *FastDTW*, la méthode *X-MeansTS* reste néanmoins, moins performante.

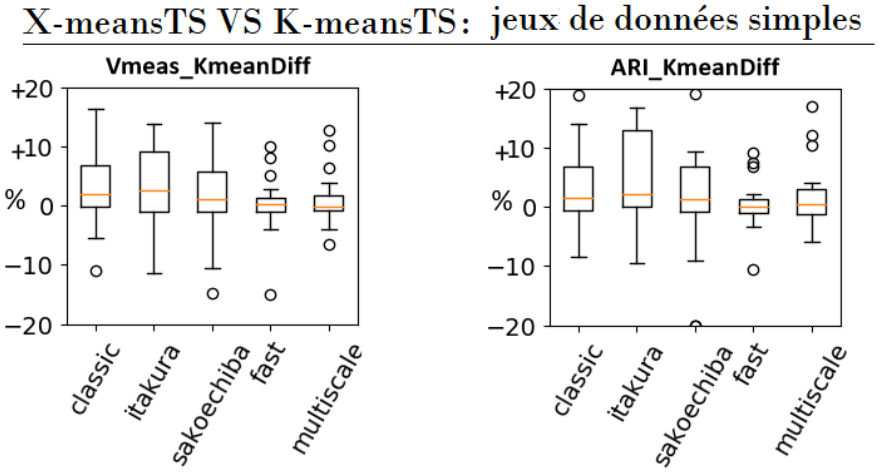

Fig. 3.19 *Vmeas_Kmeansdiff* et *ARI_Kmeansdiff* en fonction de l'ensemble de données non complexes

Comparatif de la qualité des clusters entre *X-meansTS* vs *K-meansDTW* en fonction des jeux de données particuliers : La comparaison de qualité des clustering a été faite sur les jeux de données 'particuliers'. La figure 3.20 montre, comme pour les jeux de données 'simples', les distributions de différence de qualité ($Vmeas\_Kmeansdiff$ et $ARI\_Kmeansdiff$) entre les méthodes *X-meansTS* et *K-meansDTW* et en fonction des différentes mesures dérivées de *DTW*. Les boxplots de la figure montre, comme précedemment sur les jeux de données simples, en abscisse, les différentes mesures dérivées de *DTW* et en ordonnée les distributions de taux de $Vmeas\_Kmeansdiff$

pour les boxplots de gauche et de taux de $ARI\_Kmeansdiff$ pour les boxplots de droites. La figure 3.20 montre que *X-MeansTS* crée en moyenne des clusters de meilleurs qualités que *K-MeansDTW* sur les jeux ces données 'particuliers'. Comme pour les jeux de données simples, d'après cette figure 3.20, *X-MeansTS*,selons les taux de *Vmeas_KMeansDiff*, est en moyenne 2% plus homogènes que *K-MeansDTW* sur les jeux de données simples. Les mesures de distance *DTW* combinées avec le parallélogramme de *Sakoechiba* et l'approche *Fast* ont des distributions des valeurs, là encore beaucoup plus étalées.

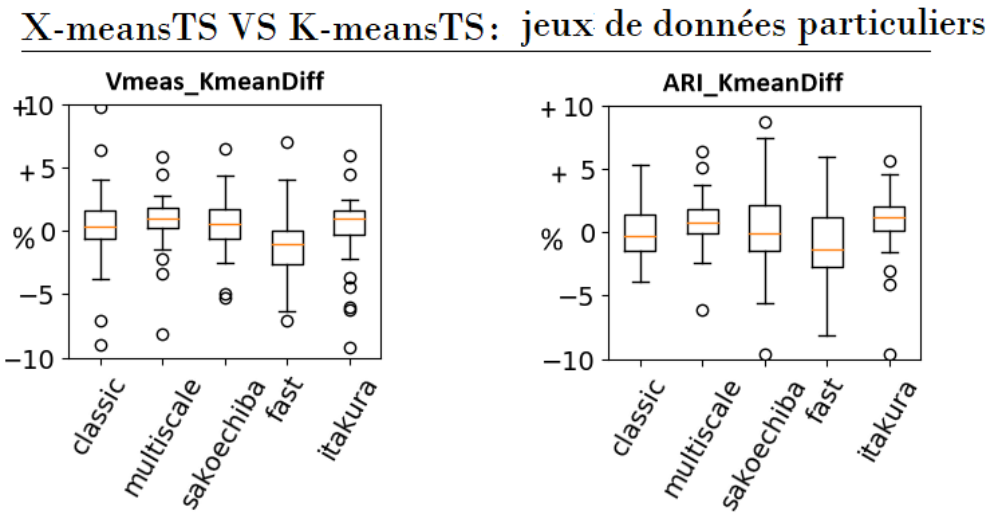

Fig. 3.20 *Vmeas_Kmeansdiff* et *ARI_Kmeansdiff* en fonction de l'ensemble de données non complexes

Comparatif de la qualité des clusters entre *X-meansTS* vs *K-Shape* sur les jeux de données simples : Comme pour le comparatif de la qualité des résultats entre *X-meansTS* et *K-meansDTW*, la même comparaison a été faite entre les clusters obtenus par *X-meansTS* et *K-Shape*. Contrairement aux scénarios de test précédent où les mesures dérivées de *DTW* été utilisées par *X-meansTS* et *K-meansDTW*, ici la comparaison est faite uniquement en modifiant la mesure de distance *DTW* en entrée de *X-meansTS*. En effet avec l'approche *K-shape* la mesure de *DTW* ne peut être intégrée puisque le séries sont redéfinies par la transformée de Fourier [128]. La figure 3.21 montre ainsi la distribution des taux de **Vmeas_ShapeDiff** (graphe de gauche) et **ARI_ShapeDiff** (graphe de droite). En abscisse les différentes mesures dérivées de *DTW*. D'après les résultats montrés sur cette figure 3.21, sur les jeux de données simples, *X-MeansTS* génère sur l'ensemble des test, des clusters (légèrement) moins homogènes et moins similaires (aux classes données réelles) que les clusters générés par *K-Shape*. En effet, on voit que les mesures de **V-mesure** et **ARI** obtenus par *K-Shape* sont en moyenne 1 à 2% plus élevées que celles de *X-MeansTS* sur la base d'un même nombre de clusters

générés par les deux méthodes (cf $ARI\_ShapeDiff$ et $Vmeas\_Shapediff$ dans la figure 3.21). On remarque que les mesures de distance *DTW* combinées avec le parallélogramme de *Sakoechiba* et la bande *Itakura* [73] ont des distributions de valeurs de performances beaucoup plus étalées.

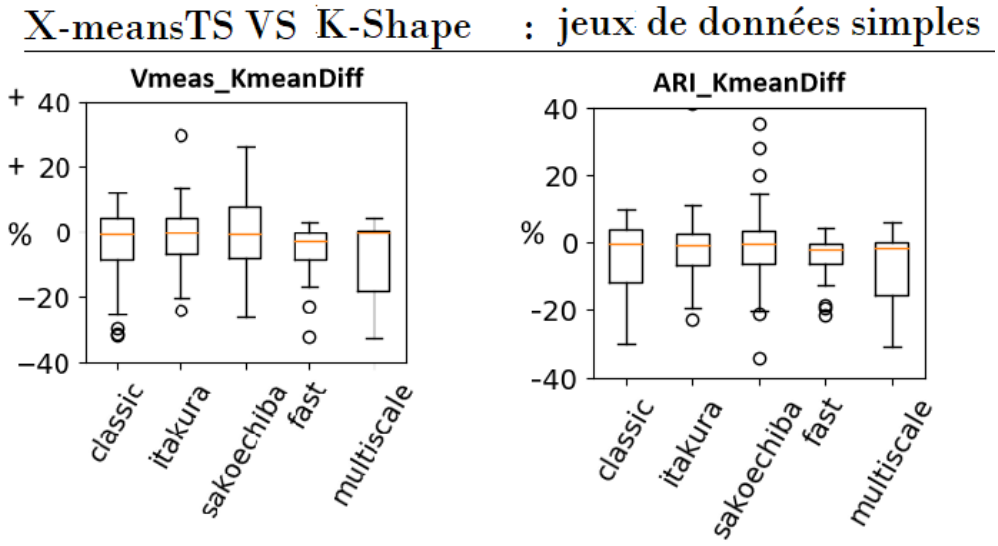

Fig. 3.21 *Vmeas_Kmeansdiff* et *ARI_Kmeansdiff* en fonction de l'ensemble de données non complexes

Comparatif de la qualité des clusters entre *X-meansTS* vs *K-Shape* sur les jeux de données particuliers : La figure 3.22 montre comme pour les jeux de données simples, la distribution des différences de valeurs des mesures de *ARI* et de *V-mesure*. La figure montre que notre méthode *X-MeansTS* est dans la majorité des résultats génère des clusters de meilleurs qualités que *K-Shape* sur les jeux de données particuliers dans lesquels le nombre de classes est très important. En effet, on voit sur la figure 3.22 que les mesures de **V-mesure** et **ARI** obtenus par *X-MeansTS* sont en moyenne 10% plus élevées que celles de *K-Shape* sur la base d'un même nombre de clusters générés par les deux méthodes (cf $ARI\_ShapeDiff$ et $Vmeas\_Shapediff$ dans la figure 3.21). *K-Shape* a de bons résultats sur les données simples mais échoue sur les jeux données particuliers.

NB : Dans le cadre de jeux de données de qualité (jeux de données équilibrées selon des labels corrélés aux attributs....) le rapport entre l'écart-type et l'entropie utilisé dans la nouvelle mesure de dispersion, est un bon indicateur d'homogénéité des clusters. En effet, la mesure de dispersion permet de sélectionner des individus répartis autour

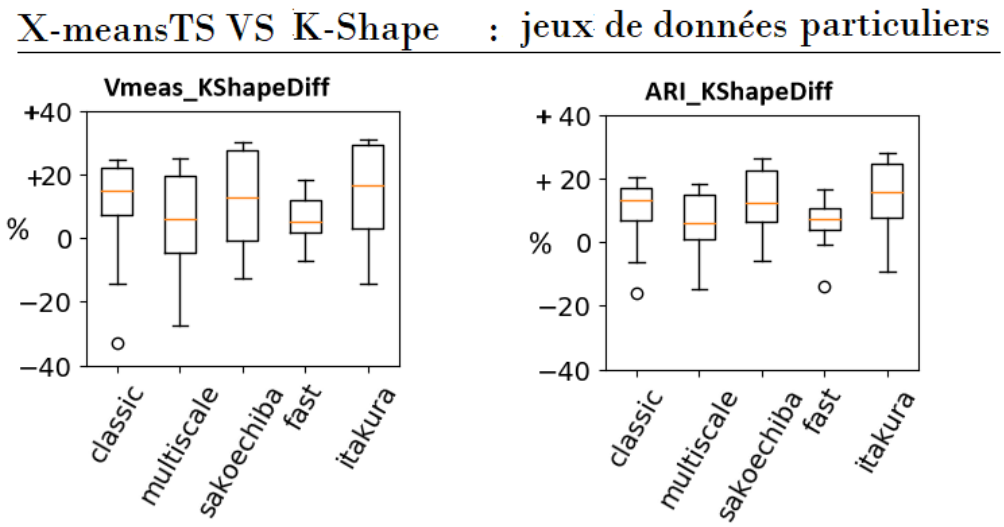

Fig. 3.22 *Vmeas_ShapeDiff* et *ARI_ShapeDiff* en fonction de l'ensemble de données non complexes

du représentant et elle s'intéresse pour cela à la probabilité d'apparition des valeurs de distances. Des séries peuvent avoir les mêmes distances avec leur représentant, lorsqu'elles ne présentent pas d'importantes variations. En cherchant à réduire l'écart type de ces distances, l'approche permet également, dans le cas des jeux de données de qualité, d'augmenter l'homogénéité. La méthode *X-meansTS* obtient des clusters plus homogènes sur des jeux de données et d'autant plus lorsque le nombre de classes est élevé. Le calcul du représentant par cluster est déterminant, pour l'homogénéité intra-cluster, et l'homogénéité globale du clutering. Dans le cas des testes, le centroïde a été le seul représentant à défaut. Notons néanmoins que les mesures de distance utilisées, déterminent un représentant qui peut être différents. Cette différence découle de l'espace de recherche du chemin optimal entre deux séries. Le chemin optimal, entre deux séries, peut être différent en fonction des différentes méthodes dérivées de la méthode *DTW*.

3.6.3 Synthèse sur les résultats expérimentaux

Les résultats précédent ont prouvé que :

- la méthode *Xmeans-TS* est plus performante que la méthode *K-Shape* sur les jeux 'particulier', et la différence de qualité des clusters obtenus sur ces jeux, est significative; Les jeux simples étant plus important, les résultats précédents ont mis en avant ce déséquilibre, en relevant des valeurs de *ARI* et de *V-Mesure* supérieur pour K-shape (qui seraient liés principalement aux jeux simples),
- la méthode *Xmeans-TS* sera plus performante que la méthode *K-meansTS*, davan-

tage sur les données simples. De plus rappelons que lorsque l'homogénéité des clusters de la méthode Kmeans-TS est plus importante, l'étendue des différences d'homogénéité (des valeurs *ARI_KMeansDiff* et *Vmeas_KMeansDiff*) l'est moins, que lorsque *Xmeans-TS* présente des clusters de meilleurs qualités.

Interprétation des mesures de distances dérivées de *DTW*

L'interprétation des performances en fonction des mesures de distance des séries temporelles peut être complexe dans certains cas. En effet, dans [54] l'impact du parallélogramme d'*Itakura* sur la précision du clustering a été comparé à celui de la bande de *Sakoe-Chiba*, et de la mesure classic de *DTW*. Dans [54], la génération des clusters en utilisant DTW avec la contrainte par *Itakura*, obtient plus souvent de meilleurs performances. Or cette contrainte (figure 3.4), impose une recherche d'un chemin optimal, dans un espace plus restreint que la méthode *Sakoe-Chiba* [73]. L'approche *Sakoe-Chiba* devrait avoir de meilleurs résultats. Toutefois, en fonction des variations des séries, la contrainte d'*Itakura*, peut conduire à un regroupement des séries différents de celui de *Sakoe-Chiba* et par conséquent, les deux mesures peuvent créer des représentants différents. En réduisant l'espace de recherche, les groupes de séries, résultant du clustering, pourraient être plus affinés.De la même manière, l'approche *Fast* [150], peut être plus ou moins performante en fonction des données. La recherche dichotomique du chemin optimal, qui est faite dans l'approche *Fast*, n'aboutie pas forcément à la recherche d'un chemin optimal, il y a une perte potentiel d'informations car la méthode débute la recherche du chemin optimal entre deux séries en réduisant la résolution des séries (cf section 3.2).

3.7 Une nouvelle approche de clustering de séries temporelles multivariées

Dans cette section, nous présenterons une nouvelle approche de clustering de séries temporelles multi-variées en s'appuyant sur la méthode *X-MeansTS*.

Dans la littérature, l'analyse des séries temporelles multivariées est utilisée lorsque l'on souhaite modéliser et expliquer les interactions et les co-variations entre un groupe de variables de séries temporelles. Considérons *n* variables de séries temporelles (s_1,..., sn). Une série temporelle multi-variée est un vecteur de *n* série temporelle s_t où la $i^{ème}$ ligne de s_t est s_{it}. Autrement dit, pour tout instant *t*, $s_t = (s_{1t}, ..., s_{nt})$. Le clustering de séries temporelles multi-variées est définie comme le regroupement du vecteur de séries temporelles en fonction des valeurs des différentes variables prises aux différents instants [44]. Notons que les approches selon la littérature passe par une réduction de la dimensionnalité des données [107, 188]. La réduction de la dimension est faite en cherchant à perdre le moins d'informations possibles. Des méthodes concernant la statistique multi-variée telle que l'analyse en composantes principales (ACP) [178], est régulièrement utilisée pour déterminer les dimensions qui expliquent le mieux la dis-

persion des données. Dans l'*ACP*, les données (par variables) sont projetées sur un axe, qui maximise la dispersion des points. Les dimensions conservées sont celles dont la projection des points sur l'axe, respectera au mieux la forme initial du nuage de points, et notamment l'inertie du nuage de ces points. Afin d'utiliser la méthode d'*ACP* pour le clustering des séries temporelles multi-variées, les séries peuvent être transformées en données statiques qui représentent des caractéristiques dérivées des séries (moyenne, écart-type...). L'*ACP* est appliquée sur ces caractéristiques afin d'identifier les deux principales dimensions qui respecte au mieux l'inertie des points définie par ces caractéristiques (dérivées des séries). Cela permettra de réduire la dimenssionalité des séries et d'utiliser les principales dimensions pour déterminer un modèle des séries.

La nouvelle approche que nous proposons s'intéresse au clustering d'ensemble de séries temporelles multivariées, et non uniquement à des séries temporelles multivariées i.e au vecteur (précité) de séries temporelles. Elle ne s'apparente pas aux approches basées sur un modèle (tel que ARMA), et qui est souvent utilisée pour le clustering de séries temporelles multi-variées. Notre méthode a une approche principalement basée sur les formes, sans réduire la dimensionnalité des données dans un premier temps. Elle utilise dans un second temps l'approche basée sur les caractéristiques. Ces caractéristiques seront les distances entres les séries temporelles multi-variées et de nouveaux descripteurs. Pour obtenir ces descripteurs (attributs), elle se sert de la méthode *X-meansTS* et génère des clusters par variable. Elle utilise les représentants, des clusters générés, comme nouveaux attributs pour le clustering des séries multi-variées. Pour cela les données sont transformées en un tableau de données statiques qui sera détaillé ensuite et dans lequel, les valeurs sont les distances (*DTW*) entre les séries et chaque représentant des variables. Nous verrons ensuite comment la méthode K-means sera appliquée sur le tableau de manière hiérarchique sur chaque cluster qu'elle génère. Il n'existe à ce jour aucune approche de clustering d'ensemble de séries temporelles multivariées, et multi-échelles, et qui utilise une approche basée sur les formes. a développé une méthode de clustering de séries temporelles multivariées, qui n'est pas multiéchelles.

[188] propose une définition d'un ensemble de séries temporelles multi-variées *S* :

1. *S* constitué de *N* série temporelles multi-variées *MTS* représentés par $S = \{ S^1$, S^2, ..., $S^N\}$.

2. Chacun des $S^i = \{1, ..., N\}$ se compose de *n* composante d'une série temporelle univariée (*CUVTS*) qui peuvent être représentés comme S^i_j, j = 1, ..., n, tel que $S^i = \{S^i_1, S^i_2, ..., S^i_n\}$.

3. Chacun des *N CUVTS* d'une *MTS*, $Sij, j = 1, ..., n$, de S^i représente une série temporelle de valeurs de données collectées sur une période de temps pour les variables, $V_j; j = 1; ...; n$, respectivement.

4. Comme $V_j, j \in \{1; ...; n\}$; est surveillé dans le temps, les valeurs que V_j prend aux instants de 1 ; ... ; p_{ij} ont été représentées par S^i_j = ($s^i_{j,1}$; $s^i_{j,2}$; ... ; $s^i_{j,t-\tau^i}$, ... , $s^i_{j,1}$, ...$s^i_{j,1+P_{ij}}$), $1 \leq \tau < P_{ij}$, $\tau \in \mathbb{Z}$ et $t = 1; ...; P_{ij}$.

L'une des principales tâches de l'analyse *MUTSCA* (Multivariate Time Series Clustering Algorithm) est de découvrir des interrelations temporelles entre des variables temporelles. Cela lui permet d'extraire des cacactéristiques. *MUTSCA* utilise ensuite l'algorithme *K-means* sur ces caractéristiques. Comme chacune des variables génère une série temporelle univariée (*CUVTS*) d'une série temporelles multi-variées (*MTS*), la tâche principale de *MUTSCA* est de découvrir les interrelations temporelles entre les valeurs observées à différents instants dans une *CUVTS* ou entre deux ou plusieurs *CUVTS*. Ces interrelations temporelles constituent les modèles temporels intra-CUVTS et inter-CUVTS respectivement dans chaque *MTS*. Pour cela, l'algorithme *MUTSCA* recherche si une variable à un instant *i* est liée temporellement à une autre valeur, $S^i_{j,t-\tau}$, précédemment observée à l'instant $t - \tau$, $1 \leq \tau < t$. *MUTSCA* détermine la différence entre la probabilité conditionnelle, $P(S^i_{j,t}|S^i_{j,t-\tau})$, et la probabilité a priori $P(S^i_{j,t})$ (appelé *lift*). Plus la différence est grande, plus S^i_j est lié temporellement à $S^i_{j,t-\tau}$. Les différences de probabilités entre $S^i_{j,t}$ et les valeurs précédemment observées, $S^i_{j,t-\tau}$ (s smax, où smax est le décalage temporel maximal qu'un utilisateur choisit d'explorer) dans S^i_j, constituent donc les modèles intra-CUVTS de S^i (Cf algorithme 10). Après avoir calculé ces modèles intra-CUVTS de S^i, l'algorithme *MUTSCA* recherche des règles significatives entre variables, i.e recherche si $S^i_{j',t-\tau}$ est toujours précédé aux positions $\tau \geq 0$ par $S^i_{j,t}$, l'algorithme en conclue que $S^i_{j',t-\tau}$. dépend de $S^i_{j,t}$. Dans ce cas, l'algorithme considère que $S^i_{j,t}$ est lié temporellement à $S^i_{j',t-\tau}$. L'ensemble de toutes les valeurs, $S^i_{j',t-\tau}$, $1 \leq \tau < t$, auxquelles $S^i_{j,t}$. est lié temporellement, constitue les modèles inter-CUVTS de S^i. Les *lifts* obtenus à partir de ces modèles sont les caractéristiques sur lesquels l'algorithme standard *K-means* est ensuite appliqué.

La qualité des cluster de l'algorithme *MUTSCA* ont été comparés avec les modèles *ARMA* et de *Markov* (sans précision sur les choix de paramètres des modèles). La précision [119] des clusters de *MUTSCA* est calculée à partir de mesure de qualité des clusters; La qualité des clusters est basée sur l'homogénéité des labels, en considérant le nombre d'enregistrements avec un label *l* et le nombre d'enregistrements total par cluster. Cet précision est réduite à environ 50% avec le modèle *ARMA* et est de 72% en utilisant le modèle de *Markov* [42].

Nous proposons dans la section 3.7.1 ci-dessous des notations et des définitions concernant l'ensemble de séries temporelles multi-variées, multi-échelles et le descriptif détaillé de notre nouvelle méthode pour le clustering de cet ensemble de séries.

Algorithm 10 MUTSCA

Input:
- $S = (S^1, S^2, ..., S^N)$,variables $S^i = (S_1^i, S_1^i, ..., S_n^i)$,

Output:
- un ensemble de *lift* pour chaque séries temporelles multivariées *(MTS)*

```
1: for chaque séries multivariées do
2:     Discrétiser chaque variable en fréquence égale pour S^i
3:     for Pour chaque variable v_1 dans MTS do
4:         obtenir des modèles intra-CUVTS
5:         for  Pour chaque variable v_2 différent de v_1  do
6:             modèles = modèles inter-CUVTS
7:             Résultat += modèles
8:             (Le résultat est un ensemble de Lift MTS)
9:         end for
10:        Résultat += inter-modèles
11:    end for
12:    finalResult += Résultat
13: end for
14: Application de Kmeans sur finalResult
```

3.7.1 Notations et définitions

Considérons un espace multidimensionnel où chaque individu est décrit par un ensemble de variables temporelles (des séries temporelles). On note $I = \{I_1, I_2, \ldots, I_n\}$ l'ensemble d'individus et $S = \{s_1, s_2, \ldots, s_m\}$ ensemble des variables temporelles défini sur un ensemble d'estampilles temporelles $\Gamma = \{T^1, T^2, \ldots, T^m\}$. Chaque estampille temporelle est associé à une variable. Deux variables s^i et s^j peuvent avoir des estampilles temporelles différentes avec une échelles différentes, c'est-à-dire, $T^i = \{t_1^i, t_2^i, \ldots, t_{p_i}^i\}$, et $T^j = \{t_1^j, t_2^j, \ldots, t_{p_j}^j\}$ et $p_i \neq p_j$.

On notera $I_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}$ ou $s_{ij} = \{s_{ij}(t_1^i), s_{ij}(t_2^i), \ldots, s_{ij}(t_p^i)\}$ une série temporelle de longueur p où $s_{ij}(t)$ correspond à la valeur du signal liée à la variable s_j de l'individu i au temps t, avec $j \in \{1, 2, \ldots, m\}$. On parle de séries temporelles multi-variées et multi-échelles (***MMTS***).

Definition 3.7.1 (**Clustering de *MMTS* et leurs représentants**). *On appelle k-clustering de MMTS de $S = \{s_1, s_2, \ldots, s_m\}$, l'ensemble $CM = \{CM_1, CM_2, \ldots, CM_k\}$ les k sous-ensembles multivariés de S avec $CM_i = \{C_i^1, C_i^2, \ldots, C_i^m\}$. CM contient k sous-ensembles de S (au sens d'une mesure de distance $Dist$), chacun ayant un représentant noté $R_{C_i^j}$ avec $\forall i \in \{1, \ldots, k\}$, $\forall j \in \{1, \ldots, M\}$ $C_i^j = \{s_{i_1 j}, s_{i_2 j}, \ldots, s_{i_{n_i} j}\}$ vérifie les critères suivants :*

1. $S = \cup_{i=1}^{k} CM_i$ *et* $CM_h \cap CM_i = \emptyset$ $\forall h \neq i$.
2. $\forall s \in C_i^j$ $Dist(R_{C_i^j}, s) < Dist(R_{C_h^j}, s)$ *avec* $h \neq i$ $\forall h, i \in \{1, \ldots, k\}$, $\forall j \in \{1, \ldots, m\}$.

3.7.2 Principe de la méthode MMTS

Le principe de l'approche que nous proposons pour clusteriser des séries temporelles multivariées et multi-échelles repose sur note méthode présentée en section 3.5.3. Notre méthode nécessite de fixer au moins les mêmes paramètres que la méthode *X-MeansTS* à savoir : .

1. nb_min_clust : le nombre de clusters initialement généré par *X-meansTS*.
2. nb_min_inst : le nombre minimum d'instances admises par cluster.
3. s_d^j : le seuil minimum de la mesure de dispersion accepté pour chaque variable j.

Notre approche de clustering *MMTS* prend en entrée les paramètres de la méthode *X-meanTS* avec des seuils de dispersion s_d^j à fixer pour chacune des variables s_j. Ce seuil ne prend pas la même valeur pour chaque variable. Le nombre de clusters choisi pour le découpage initial et le nombre d'instances minimales, sont quant à eux, paramétrés de manière égale pour les clusters de chaque variable.

La figure 5.3, présente le principe de l'approche *MMTS*. Dans la première étape, l'algorithme *X-meansTS* est appliqué à chaque variable indépendamment. Cette étape fournit un ensemble de clusters pour chaque variable. Le nombre de clusters peut être différent par variable. Les paramètres liés au nombre minimal d'individus (nb_min_inst) et au nombre de clusters initiaux (nb_min_clust) sont identiques pour chaque variable lors de l'appel de *X-meansTS*. Le seul paramètre qui diffère par variable est le seuil de dispersion. Pour chacune d'elle, comme pour les résultats présentés de la méthode *X-meansTS*, un seuil moyen est recherché, en fonction d'un nombre de clusters (par variable), générés sans appliqués la mesure de dispersions $disp$. On calcule la dispersion par cluster par variable. Ensuite la valeur moyenne ou maximale est retenue, qui es donc différente par variable. Cette valeur sera la valeur du seuil de dispersion, lorsque *X-meansTS* est appelé (dans *Xmeans-MMTS*) sur les séries d'une variable donnée. Les scénarios de générations des paramètres seront détaillés ensuite dans la section 3.7.4

Dans la seconde étape de la méthode *X-meansTS*, nous transformons les données séries temporelles en données statiques en construisant une matrice *M* ($M = (m_{ij})$) de dimension $N \times L$ (où N est le nombre d'individus, L est le nombre total de clusters fournis à la sortie de la première étape). La matrice M représente en ligne les individus et en colonne les représentants des clusters pour chaque variable. Le coefficient m_{ji} représente la distance entre chaque série de l'individu i par variable au représentant des clusters (pour la même variable). En d'autre terme si, pour chaque variable s_j, $X - MeansTS$ génère c_j clusters alors le nombre de colonne $L = \sum_{j=1}^{m} c_j$. La troisième étape considère la matrice M comme un nouveau jeu de données statique

dont les nouvelles caractéristiques représentent les distances de chaque série d'un individu aux représentants des clusters en sortie de la première étape (ceci se fait par variable). Sur cette matrice, nous appliquons une nouvelle stratégie de clustering basée sur un découpage hiérarchique et récursif et respectant le critère d'homogénéité lié à notre mesure de dispersion $disp$.

Le découpage hiérarchique et récursif, comme le montre la figure 5.3 pour l'exemple donné avec $nb\_min\_clust = 2$, procède au partitionnement des individus à partir de la matrice M. Initialement, le partitionnement est généré par la méthode classique *k-Means*. Ensuite, nous procédons au découpage itératif de chaque cluster en nouveaux clusters tant que le critère de dispersion n'est pas vérifié pour au moins un cluster d'une des variables séries temporelles. Le découpage s'arrête lorsque, tous les clusters et pour toutes les variables, la dispersion de chaque cluster est inférieure au seuil fixé pour la variable concernée. Le découpage itératif est réalisé par l'application de la méthode *K-Means* sur la matrice *M* réduite aux individus du cluster à re-découper. En revanche, le critère itératif est basé sur la mesure de dispersion appliquée aux données d'origines (séries temporelles) des individus concernés.

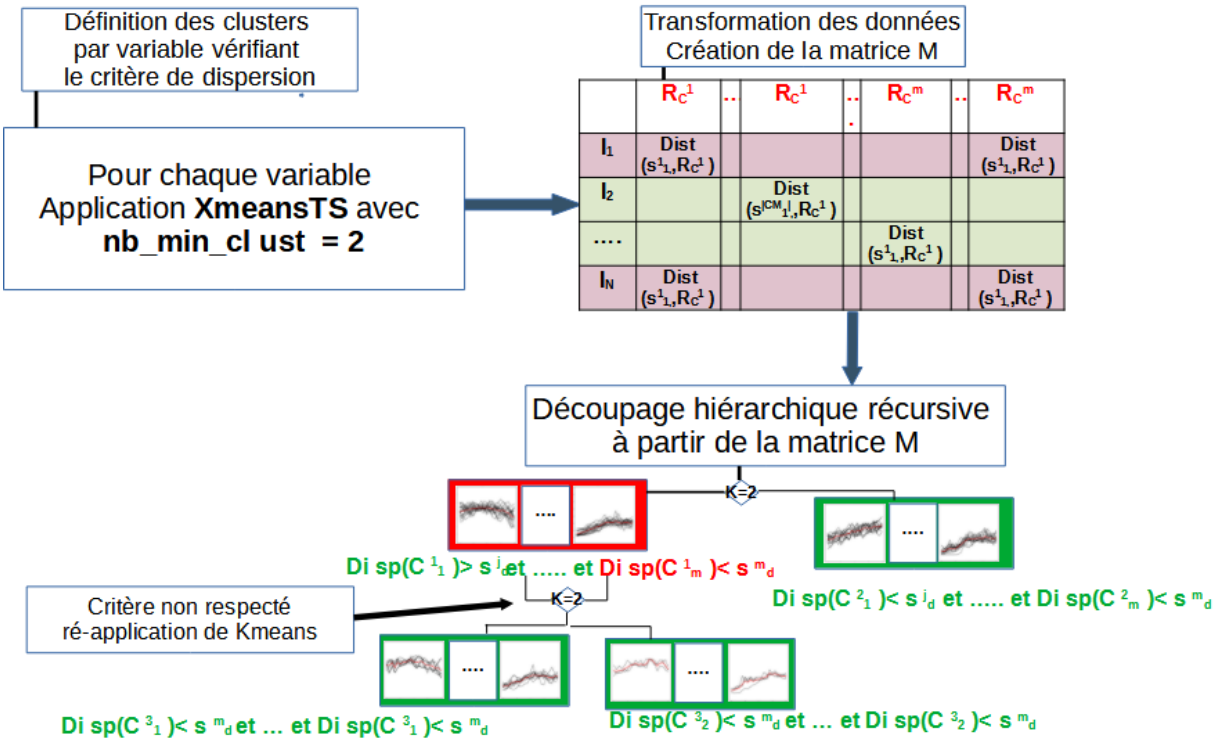

Fig. 3.23 principe de la méthode de clustering multi-varié *Xmeans-MTS*

L'approche proposée génère en plus un nombre de clusters qui est déterminé automatiquement selon les 3 critères énumérés précédemment. Nous rappelons néanmoins que l'algorithme *X-meansTS* détermine automatiquement un seuil minimum pour cha-

cune des variables, si aucun seuil n'est paramétré.

Les étapes suivantes, détaillent le déroulement du nouvel algorithme que l'on appellera *X-MeansMMTS* :

1. **Étape initiale (Génération d'un clustering par variable) :** Pour chaque variable j, les instances de I (ensemble de séries temporelles $\{s_{1j}, s_{2j} \ldots, s_{nj}\}$ de la variable j) sont partitionnées en un nombre k_j de clusters ($\{C_1^j, C_2^j, \ldots, C_{k_j}^j\}$) vérifiant l'homogénéité au sens de notre mesure de dispersion ($disp$). On applique notre algorithme *X-MeansTS* avec une mesure de distance $Dist$ (qui peut être DTW, etc.). Le seuil minimum que doit vérifier la mesure de dispersion pour chaque variable j sera noté s_d^j (Cf algorithme 9 *X-meanTS*).

2. **Etape 2 (Transformation des données) :** Après l'application de *X-meansTS* par variable, les représentants de chaque cluster i et pour chaque variable j, noté $R_{C_i^j}$ sont calculés, $\forall j \in \{1, 2, \ldots, m\}$ et $\forall i \in \{1, .., k_j\}$. Ces représentants seront utilisés pour construire de nouveaux attributs statiques sur lesquels on pourra appliquer une méthode de clustering classique. En effet, les données, représentant l'ensemble des séries temporelles multi-variées, sont transformées en données statiques représentées par une matrice notée $M = (m_{ij})$ (où $i = 1, .., n$ et $j = 1, .., L$ avec $L = \sum_{l=1}^{m} k_l$: nombre total de clusters toute variable confondue) décrivant les individus par de nouveaux descripteurs. La valeur $m_{ij} = Dist(s_{ij}, R_{C_l^j})$, pour $i = 1, .., n$ et pour tout $j = 1, .., m$, $l = 1, .., k_j$. Le tableau 3.2 montre cette transformation et la structure de la matrice M. Ainsi pour un individu i et une variable j, la matrice représente les distances entre la série temporelle de l'individu i pour la variable j à tous les représentants des clusters liés à la variable j.

3. **Etape 3 (Découpage hiérarchique récursif à partir de la matrice M):** Dans cette étape nous appliquons un découpage récursif à partir d'un clustering de base appliqué initialement à la matrice M avec un nombre minimum de clusters (nb_min_clust). Le découpage, basée sur le critère de l'homogénéité selon notre mesure de dispersion, est effectué sur les clusters en considérant l'ensemble des séries temporelles. Le principe du découpage se fait de manière récursif tant qu'il existe un cluster d'une des variables séries temporelles ne vérifiant pas le critère d'homogénéité. Pour générer les clusters initiaux à partir de la matrice M, nous utiliserons la méthode classique $k - Means$ avec $k = nb\_min\_clust$.

   (a) **Critère de découpage hiérarchique** : Le critère de découpage est basé sur la mesure de dispersion par cluster. Cette mesure à une étape de découpage hiérarchique est recalculée en fonction du représentant du nouveau cluster et ses individus.

	$R_{C_1^1}$	...	$R_{C_{k_1}^1}$	...	$R_{C_1^m}$	...	$R_{C_{k_m}^m}$
I_1	$Dist(s_{11}, R_{C_1^1})$		$Dist(s_{11}, R_{C_{k_1}^1})$		$Dist(s_{1m}, R_{C_1^m})$		$Dist(s_{1m}, R_{C_{k_m}^m})$
I_2	$Dist(s_{21}, R_{C_1^1})$		$Dist(s_{21}, R_{C_{k_1}^1})$		$Dist(s_{2m}, R_{C_1^m})$		$Dist(s_{2m}, R_{C_{k_m}^m})$
...							
I_n	$Dist(s_{n1}, R_{C_1^1})$		$Dist(s_{n1}, R_{C_{k_1}^1})$		$Dist(s_{nm}, R_{C_1^m})$		$Dist(s_{nm}, R_{C_{k_m}^m})$

Table 3.2 Matrice M des distances entre les séries des individus et les représentants des clusters multi-variés.

(b) **Condition d'arrêt de l'appel hiérarchique de Kmeans :** Le clustering est donc effectué de manière récursive sur les individus de chaque cluster $CM_i \in CM$ ($CM_i = \{C_i^1, C_i^2, \ldots, C_i^m\}$) générés à une étape par *K-Means* tant que les séries temporelles multivariées associées à ses individus ne respectent pas les critères de dispersion pour au moins une des variables i.e , $\exists j \in \{1, \ldots, m\} / Disp(C_i^j) \geq s_d^j$. Pour ne pas générer des clusters non significatifs, nous rajoutons un autre critère d'arrêt sur le nombre minimum d'individus par clusters qui devrait être toujours supérieur à nb_min_inst (seuil fixé au préalable.

3.7.3 Expérimentation de notre approche

La nouvelle approche *X-MeansMMTS* a été expérimentée sur des jeux données (séries temporelles multivariées) de l'archive *UCR* [7]. Ce jeu de données est composé de 30 ensembles de données avec un large éventail de cas, de dimensions et de longueurs de séries.

Nous avons utilisé 7 jeux de données avec des séries de longueurs différentes, des dimensions différentes (nombre de variables). Chaque jeu de données est labélisé (avec un nombre de classes variant de 2 à 39). Les classes d'appartenance des individus par jeu de données seront utilisées pour calculer les performances de la méthode *X-MeansMMTS*, i.e l'homogénéité et la complétude du clustering.

Le tableau 3.3 présente les caractéristiques de chaque jeu de données testés.

Comme nous l'avons énoncé, aucune approche similaire à *Xmeans-MMTS* existe à ce jour, qui permet le clustering d'ensemble de séries temporelles multi-variées, multi-échelles, et qui, de plus, est disponible en ligne en langage informatique. Notons également que le développement de cet algorithme s'est fait en fin de thèse et que, par soucis de comparaison de la qualité de clustering avec une méthode existante, il n'était pas envisageable, par manque de temps, de développer la méthode *CUMVTS*. De plus la méthode *CUMVTS* est d'avantage basée sur des caractéristiques et ne s'intéresse qu'aux des intervalles de temps spécifiques des séries et non à l'ensemble des valeurs des séries.

Algorithm 11 X-meansMMTS :

Input:
- $S = \{S_1, S_2, \ldots, S_m\}$ séries temporelles multivariées
- nb_min_clust : nombre minimum de clusters à l'étape initiale
- $s_d = \{s_d^1, s_d^2, \ldots, s_d^m\}$: seuil de dispersion
- nb_min_inst nombre minimum d'instances

Output: - C_{FM} ensemble de clusters

if firstCall **then**
  for j in $\{1, 2, \ldots, M\}$ **do**
    C_F^j = X-MeansTS(S_j, nb_min_clust, s_d^j ,*nbClust*, *nbMaxIter*, *recursifCpt*)
  end for
end if
I : individus $\in S$
if $|I| > nb\_min\_inst$ **then**
  isDispOk = True
  Générer la matrice $M(x, y) = Dist(x, y)$ ou $y \in R_{C_F^h}$ et $x \in S^h$
  K-means($M()$)= $\{Cluster_1, Cluster_2, \ldots, Cluster_k\}$ avec k = nb_min_clust
  $CM = \{CM_1, CM_2, \ldots, CM_k\}$ k sous-ensembles multivariés de S avec $\forall I_i \in I, I_i \in CM_i \cap Cluster_i$
  for C_{M_i} in C_M **do**
    j = 0
    isDispOk = True
    for C_i in C_{M_i} **do**
      if $Disp(C_i) < s_d^j$ **then**
        isDispOk = False
      end if
      j = j+1
    end for
    if isDispOk == False **then**
      $S_{temp} = \{S | I \in S \cap C_i\}$ avec I les individus représentés par les séries temporelles multivariés
      X-MultiTS(S_{temp},nb_min_clust,s_d, nb_min_inst)
    else
      $C_{FM} = C_{FM} + CM_i$
    end if

  end for
end if
return C_{FM}

Pour ces raisons, la comparaison de la qualité du clustering de notre approche avec une autre approche n'est pas envisageable. Néanmoins dans les résultats, nous nous intéresserons aux mesures d'homogénéité des clusters en fonction de différents seuils de dispersion. En effet si celui ci est faible alors les clusters devraient contenir peu d'instances car pour ces cluster, les séries multi-variées auront été affinées autour du représentant et l'homogénéité devrait être élevé. Par conséquent l'homogénéité dimin-

Name	# Time Series	# Dimensions	Series Length	# Classes
ArticularyWordRecognition	275	9	144	25
Heartbeat	204	61	405	2
JapaneseVowels	270	12	29	9
Libras	180	2	45	15
NATOPS	180	24	51	6
PhonemeSpectra	3315	11	217	39
UWaveGestureLibrary	120	3	315	8

Table 3.3 Multivariate time series datasets description from the UCR repository

uerait en augmentant le seuil.

3.7.4 Scénarios de test

La méthode *Xmeans-MMTS* a été appliquée plusieurs fois sur chaque jeux de données, en modifiant le seuil de dispersion et le nombre de clusters initiaux. A chaque exécution de la méthode, un seuil différent est généré pour chaque variable. La recherche de ce seuil est faite comme dans les scénarios de test de la méthode *X-meansTS*, c'est à dire qu'un premier découpage est réalisé et le calcule de la dispersion, est effectué sur chaque cluster (par variable dans le cas de *Xmeans-MMTS*). Les valeurs moyennes et maximales des différentes valeurs sont retenues pour les testes. Cette approche est effectué, par variable, et à chaque teste de *Xmeans-MMTS*. Concernant le paramètre nb_min_clust (nombre de clusters initiaux), il est identique pour chaque variable. Pour chaque jeu de données, 5 valeurs pour le paramètre nb_min_clust sont générées en fonction du nombre réel de classes du jeu. Ces valeurs sont comprises dans l'intervalle $[clr - 2, clr + 2]$ avec *clr* : le nombre réel de classe du jeux de données. Le paramètre nb_min_inst, le nombre minimum d'instances, varie de 3 à 10 pour les tests. Ainsi, sur la base 'de ces nombres minimums d'instances' (i.e des différentes valeurs possibles du paramètre nb_min_clust) et des seuils de dispersions (minimum, médiane et maximum), *Xmeans-MMTS* est testée selon les combinaisons possibles de ces paramètres sur chaque jeu de données. Comme exemple, si le nombre réel de classe vaut 5, il y a huit valeurs de nb_min_inst, 5 valeurs différentes de nb_min_clust, et les trois valeurs de seuils de dispersion calculés. Pour chaque jeu de données, la méthode *Xmeans-MMTS* est donc testée $5 \times 3 \times 8 = 120$ fois avec différents paramètres d'entrées. La qualité des clustering sont évaluées selon la V-measure,l'homogéneité et la complétude.

Le tableau 3.4 des résultats montre les valeurs moyennes des performances obtenues avec deux seuils (minimum et médian) sur chaque ensemble de données. La colonne "#Clusters finaux" montre le nombre moyen de clusters multivariés obtenus (les valeurs ont un écart-type égal à 20). Ce nombre moyen est très proche du nombre réel de classes

Dataset	Final clusters (± 20)	V-measure	Homogeneity	Completeness	Name s_d
WordRecognition	141	0,92	0,85	1	minimum
Heartbeat	56	0,74	0,59	1	minimum
JapaneseVowels	6	0,48	0,31	1	minimum
Libras	76	0,88	0,78	1	minimum
NATOPS	86	0,89	0,8	1	minimum
PhonemeSpectra	86	0,37	0,23	1	minimum
GestureLibrary	51	0,88	0,79	1	minimum
WordRecognition	111	0,81	0,68	1	médiane
Heartbeat	26	0,38	0,23	1	médiane
JapaneseVowels	6	0,48	0,31	1	médiane
Libras	41	0,81	0,68	1	médiane
NATOPS	26	0,58	0,41	1	médiane
PhonemeSpectra	6	0,34	0,2	1	médiane
GestureLibrary	36	0,81	0,68	1	médiane

Table 3.4 Mean performance of *X-MeansMMTS* method on datasets from UCR repository

par jeu de données, obtenu automatiquement grâce à la méthode *X-MeansMMTS* avec un seuil maximal Néanmoins ce nombre de cluster final peut être éloigné du nombre réel de classe, lorsque le seuil est faible. Au final, le tableau de résultats 3.4 montre que sur des jeux de données complexes (par exemple, le jeu de données Phonème dans le tableau 3.3), la performance du *X-MeansMMTS* reste excellente, ce que la complétude étant de 1 indique également. Le seuil de la mesure de dispersion permet d'augmenter l'homogénéité des clusters multi-variés, lorsque le seuil minimum est de plus en plus faible.

3.7.5 Perspective d'amélioration de l'approche *X-MeansMMTS*

La recherche automatique du seuil de la mesure de dispersion peut être améliorée, afin d'optimiser l'homogénéité finale (de l'ensemble des clusters). Cette mesure peut être appliquée à toutes les méthodes de clustering basées sur des mesures de distance, par exemple elle peut être directement adaptée à la méthode existante *K-Means*, pour le clustering de données statiques. En effet, la méthode multivariée génère une matrice de données statiques, sur laquelle *K-Means* est appliqué de manière hiérarchique. Une des perspectives de ce travail est également de superviser l'approche *X-MeansMMTS*. Pour cela, nous pouvons nous appuyer sur des approches d'apprentissage supervisé basées sur l'exploitation des distributions gaussiennes des données. En effet, la mesure de dispersion est adaptée à ce type de distribution. Enfin, l'objectif est de créer un modèle de classification supervisée qui puisse être interprétable à partir des séries temporelles

multi-échelles et multivariées. En effet, il n'existe actuellement aucune méthode qui réponde à cette problématique.

Chapitre 4

Élevages de stylirostris sur la Grande Terre

En Nouvelle-Calédonie, la qualité de l'eau de son lagon, favorise le développement d'une bio-diversité aquatique exceptionnelle. Avec des conditions climatiques favorables au développement de filières aquacoles, cela a conduit à l'émergence de politique de développement d'une filière crevetticole soutenue par les institutions (Territoire, Provinces, État). La création de la ferme *Sodacal* (1983), avec le soutien technique, scientifique et commerciale de l'IFREMER (Institut Français de recherche pour l'exploitation de la mer), a conduit au développement de la filière en s'appuyant dans un premier temps sur la consommation locale, pour ensuite atteindre une clientèle internationale.

La filière est aujourd'hui composée de deux provendiers, de deux ateliers de transformation dont un traite la totalité des volumes exportés (La SOPAC Société des Producteurs Aquacoles Calédoniens), de quatre écloseries, de 19 fermes de grossissement installées sur la côte Ouest du pays, d'un centre technique Aquacole (CTA) et dispose de l'appui scientifique de l'Ifremer et de l'Agence Rurale (suivi économique). Les fermiers se sont regroupés dans une association le GFA (Groupement des Fermes Aquacoles). Les fermes aquacoles sont, pour la majorité, implantées sur des zones salées à l'arrière des mangroves, appelées "tannes".

4.1 Les données relevées par la filière crevetticole Calédonienne

Nous étudierons les données d'élevages provenant de fermes crevetticole Calédonienne, collectées par le *GFA* (Groupement des Fermes Aquacoles). L'étude a été conduite, (Cf figure 4.1)), sur 17 fermes positionnées du nord au sud, sur la cote ouest du pays. Elle porte sur 700 élevages de crevettes réalisés entre 2000 et 2016.

Un élevage s'effectue sur une durée de 4 à 6 mois. Afin d'assurer un suivi lors de l'élevage, différents types de données sont recueillis au cours du temps (paramètres physico-chimiques et biologiques (données zootechniques...) par les éleveurs. Notre travail intégrera également les données collectées à l'usine de conditionnement **SOPAC** (Société des Producteurs Aquacoles Calédoniens). Ces données sont essentiellement des données de qualité du produit. La qualité de chaque production d'élevage est

Fig. 4.1 Les fermes du GFA

évaluée à partir d'une analyse qualitative et quantitative en laboratoire. Elle est aussi évaluée en fonction des différents calibres des crevettes péchées. Les données d'élevages comptabilisent plus de 1.2 millions de données et celles de qualité concernent plus 200 000 mesures. Cette étude est la première étude dans le domaine aquacole portant sur une quantité de données de production et de qualité aussi importante, et collectées sur une durée significative de 17 années.

4.1.0.1 Clause de confidentialité

Les données sont soumises à une convention de confidentialité tripartite entre l'ISEA l'Institut de Recherche en Sciences Exactes et Appliquées, le GFA et la SOPAC. Selon cette convention, les fermes ne doivent pas être identifiables dans l'analyse que nous ferons des données à disposition. Les bassins et les élevages seront associés à des identifiants numériques.

4.1.1 Le processus de grossissement des crevettes et les données relevées

4.1.1.1 La ferme aquacole

La figure 4.2 montre une prise de vue aérienne d'une ferme située au nord de la Nouvelle-Calédonie. Des vignettes montrent quelques images d' équipements présents dans chaque ferme, qui permettent d'alimenter un bassin en eau. La surface des bassins varie entre 4 et 11 hectares pour une hauteur d'eau moyenne d'un mètre. L'eau est pompée directement dans le lagon, par une station de pompage (point gris dans l'image). Elle

circule vers l'entrée des bassins, au travers des digues (suivant la flèche bleu). Pour alimenter ces bassins, il y a des entrées et des sorties qu'on aperçoit en bas à droite dans l'image. Ces dispositifs servent au renouvellement de l'eau et aux pêches. Un adule se définit pas sa capacité à se reproduire. Ce qui n'est pas le cas pour les crevettes pêchées. Ce sont des juvéniles

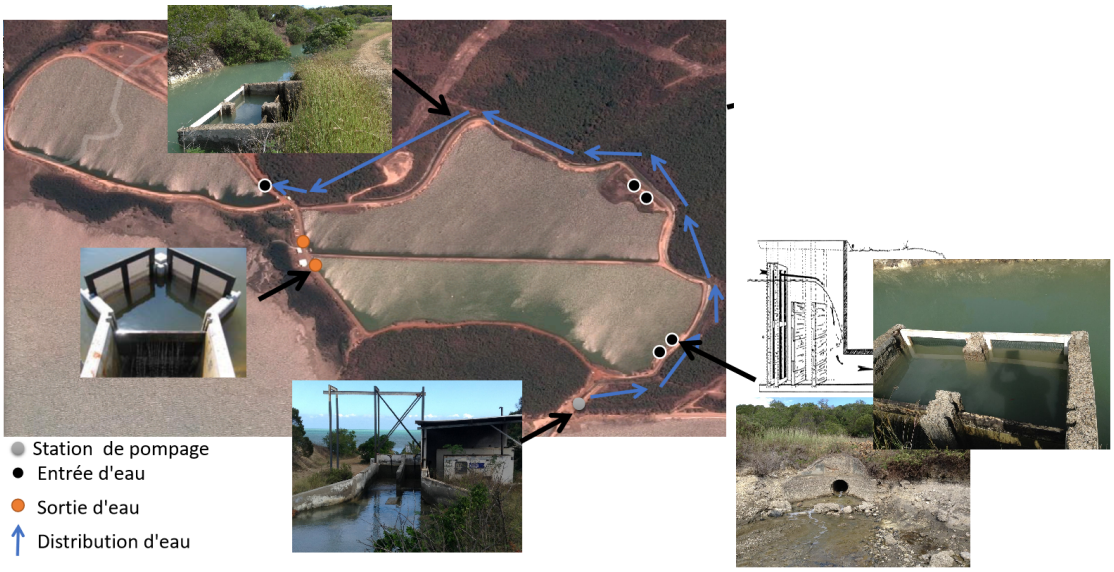

Fig. 4.2 Fonctionnement d'une ferme aquacole

Durant l'élevage, le renouvellement de l'eau des bassins, peut-être hebdomadaire dans les premières semaines d'élevage. Par la suite, la fréquence devient journalière; Le renouvellement est essentiellement un processus de dilution de la colonne d'eau qui a tendance à s'eutrophiser avec la durée de l'élevage [102].

Le processus de grossissement des crevettes peut-être décomposé en plusieurs phases décrites dans la section suivante (cf. figure 4.3). Durant chacune de ces phases, des données sont relevées de manière automatique ou suivant des protocoles d'échantillonnage manuels. Ces données sont le plus souvent des données standards de production qui assurent le suivi de l'élevage en cours.

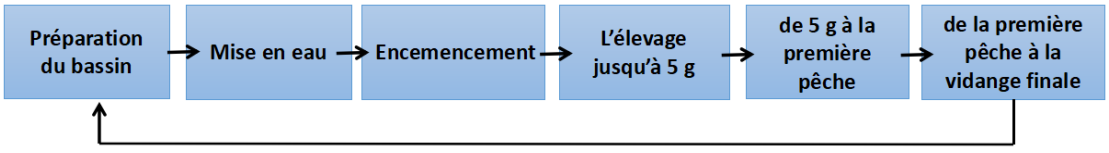

Fig. 4.3 Les phases d'un élevage

Les données relevées durant ces phases seront décrites selon leurs types dans la sous section suivante. Les données peuvent être acquises durant toute la durée d'un élevage

et dans toutes les phases de grossissement de la crevette.

4.1.1.2 La préparation du bassin

Dans le cas d'un bassin dans lequel au moins un élevage a été réalisé, la préparation du bassin débute par la phase dite d'assec. Cette phase correspond à l'assèchement du bassin afin de re-minéraliser la matière organique qui s'est accumulée suite à l'élevage antérieur (figure 4.4). Lorsque la matière s'est fortement accumulée, un labour peut-être réalisé à l'aide d'un tracteur pour travailler le fond du bassin. Le bassin est ensuite rempli avec de l'eau salée pompée dans le lagon. Différents types de données (cf. tableau 4.1) sont relevées au cours et en fin d'assec. Ce sont des données textuelles, catégorielles, et quantitatives.

Fig. 4.4 l'assec d'un bassin avant son remplissage.

Description	Unité	Quantitatif	Catégoriel	Textuel	Statique	Temporel
le niveau d'accumulation des matières organiques (sans, faible, forte)			X		X	
la durée d'assec	jours	X			X	
surface sol travaillé	%					
profondeur travaillé	cm	X			X	
pluviométrie	mm	X				X

Table 4.1 Les types de données enregistrées durant la préparation du bassin

4.1.1.3 La mise en eau

La mise en eau est l'ensemble des opérations qui créent les conditions optimales pour un bon ensemencement des post-larves de crevette. Ces post-larves font environ 3 mm. Elles sont donc très fragiles. Les conditions citées, sont liées au développement d'une

chaîne alimentaire (par exe. phytoplancton) qui servira d'alimentation aux animaux ensemencés.

Une coloration de l'eau est recherchée au cours de cette phase. Cette coloration appelée bloom, est un processus de concentration du phytoplancton dans une masse d'eau (cf. figure 4.5). Des fertilisants peuvent être utilisés pour favoriser le développement de ces organismes.

Fig. 4.5 Concentration du phytoplancton nécessaire à l'alimentation des post-larves

Il y a principalement des données quantitatives liées à la phase de mise en eau (tableau 4.2).

Description	Unité	Quantitatif	Catégoriel	Textuel	Statique	Temporel
Durée de la mise en eau	jours	X			X	
La fertilisation : par ex. Quantité d'urée (kg)			X	X		

Table 4.2 Les types de données enregistrées durant la mise en eau

4.1.1.4 L'ensemencement

L'ensemencement consiste à transférer les crevettes depuis une écloserie industrielle. Il y a 3 en Nouvelle-Calédonie (Cf. figure 4.1)). Durant le transport dans des cuves (Cf. figure 4.6), l'eau est oxygénée pour maintenir une concentration en oxygène adéquate à la survie des post-larves.

4.1.1.5 La phase de grossissement

Les différentes caractéristiques liées à la production, présentées dans le tableau 4.4, sont enregistrées au cours du temps à chaque élevage. Ces données relevées lors de la production sont principalement des séries temporelles représentant l'évolution de plusieurs variables de qualité de l'eau, par élevage. L'objectif de l'éleveur est de maintenir une

Fig. 4.6 Cuves assurant le transport des post-larves d'une écloserie vers une ferme

Description	Unité	Numérique	Catégoriel	Textuel	Statique	Temporel
Nombre de post-larves ensemencé		X			X	
Age des post-larves	jours en nurserie	X			X	
Oxygène au départ et à l'arrivée		X				X

Table 4.3 Les types de données enregistrées durant l'ensemencement

bonne qualité d'eau, essentielle dans la gestion de sa ferme pour l'obtention d'une croissance et d'une survie optimale [18]. Compte tenu de la faible profondeur des bassins ($\tilde{1}$m), il s'agit donc pour le fermier de maîtriser le degré d'enrichissement de son milieu pour éviter les crises dystrophiques préjudiciables à la santé des animaux [104].

Description	Unité	Numérique	Catégoriel	Textuel	Statique	Temporel
Température	^oC	X				X
Oxygène	$ml.l^{-1}$	X				X
fluorescence	%	X				X
Secchi	cm	X				X
pH	pH	X				X
Salinité	SP	X				X
Turbidité	NTU	X				X
Renouvellement de l'eau	taux	X				X
Alimentation	g/m^2	X				X
Poids moyen	g	X				X

Table 4.4 Les types de données enregistrées durant l'élevage

Les mesures des paramètres environnementaux sont réalisées pour évaluer les conditions d'élevage qui peuvent se révéler dans certaines conditions stressantes pour les animaux et même les affaiblir face à l'émergence de maladies [14, 97] .

L'intensité du stress augmente avec la diminution du pH de $6,5$ à $7,4$ [103]. Une concentration en oxygène inférieure à $3mg.l^{-1}$ est considérée comme stressante. Elle devient dangereuse pour des valeurs inférieures à $1mg.l^{-1}$ [98, 122]. Une température de $22°C$ est considérée comme la limite du preferendum thermique pour la crevette

élevée en Nouvelle-Calédonie [175]. L'animal devra réguler son osmorégulation pour s'adapter aux variations de salinité [101]. Ce processus nécessite une dépense énergétique supplémentaire pour les animaux [111]. Les paramètres environnementaux sont aussi des indicateurs pour évaluer la qualité du milieu d'élevage et plus spécifiquement le niveau d'eutrophisation. La fluorescence est un proxy de la biomasse phytoplanctonique. Le secchi permet d'évaluer la transparence / la turbidité d'une colonne d'eau. Il consiste en un disque d'une vingtaine de centimètres, partagé en quarts alternés noirs et blancs. Une eau avec un secchi élevé et/ou une turbidité trop faible est considérée comme défavorable aux élevages. La variation journalière de l'oxygène permet aussi d'évaluer indirectement l'enrichissement du milieu en phytoplancton [102].

La gestion de la qualité de l'eau par l'éleveur, se fait essentiellement par renouvellement de l'eau. Le volume échangé permet de limiter l'accumulation de résidus métaboliques issue de la consommation d'aliment (dont certains composés toxiques) et d'évacuer les déchets organiques permet aussi de réguler les paramètres physico-chimiques vitales et notamment l'oxygène. Le système de pompage détermine la capacité horaire de renouvellement d'eau. Un manque de renouvellement d'eau peut induire des stress (principalement des crises d'oxygène), mais trop d'apport d'eau peut déstabiliser la productivité naturelle et provoquer une turbidité insuffisante [63].

Les quantités d'aliment distribuées pendant la phase de démarrage des bassins dépendent de la saison (température) et de leur richesse en production naturelle. Divers types de granulés sont utilisés en fonction de la période de grossissement. Des fines sont privilégiées jusqu'à ce que la crevette atteigne $0,5g$ ensuite du concassé, et enfin des granulés entier dès que la crevette atteint les $2g$.

4.1.1.6 De la première pêche à la vidange

Des pêches sont réalisées à partir du 120ème jour d'élevage lorsque le poids moyen atteint environ $20g$, afin de retirer une partie de la production de crevettes. Cela permet de limiter les risques de perte de biomasse liée à des crises dystrophiques qui conduisent à une chute de la concentration en oxygène préjudiciable à la survie des animaux élevés. Ces pêches permettent également de faciliter la gestion des ressources, par une gestion de la biomasse en élevage. Une diminution de la biomasse permet d'augmenter le poids des animaux pour les pêches suivantes. Le prix des animaux augmente avec leur poids.

4.1.2 La base de données étudiées

Les données d'élevages proviennent de la base de données *STYLIBASE*, qui est une base de données relationnelle normalisée, créée par l'Ifremer et gérée depuis 2011 par le GFA [156, 137]. *STYLIBASE* a pour vocation de rassembler au sein d'une même base fonctionnelle, les données issues des bases unitaires (Stylog Module Ferme), alimentées

indépendamment par des fermes Calédoniennes.

Stylog Module Ferme (SMF) est une base *MS Access* utilisée sur les fermes pour saisir les données d'élevage. Il y a une base de données *SMF* par ferme pour une saisie locale de l'information et une gestion autonome des données. Les données de qualité proviennent d'un fichier sous format Excel fournit par la SOPAC.

La figure 4.7 montre que STYLIBASE est une base de données complexes, possédant des tables avec des relations complexes. Afin de simplifier le schéma relationnel com-

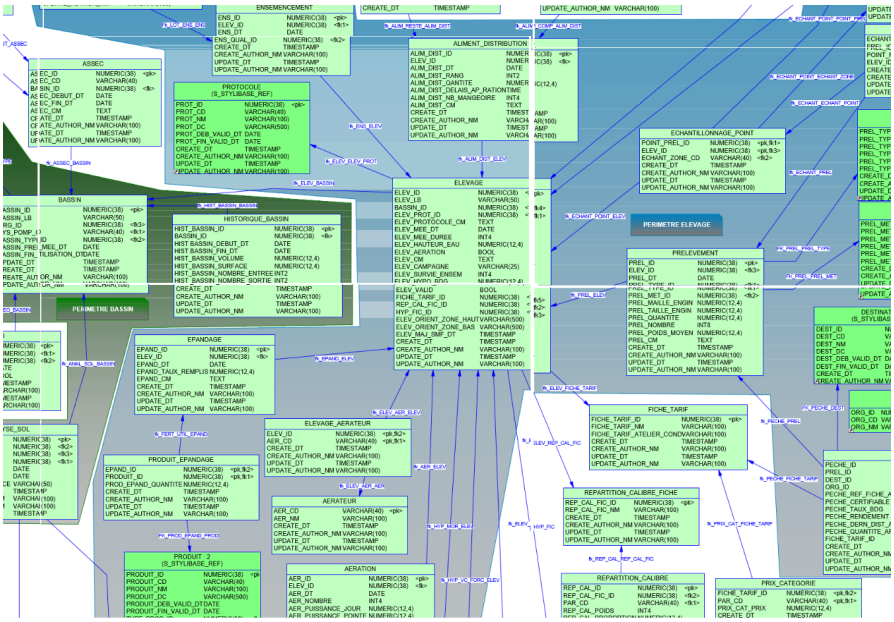

Fig. 4.7 Schéma relationnel de la base de données STYLIBASE

plexe de cette base, nous représenterons les relations entre les données étudiées selon le schéma de la figure 4.8. Dans ce schéma les données de qualité production, relevées par l'usine la *SOPAC*, sont logiquement intégrées.

La figure 4.8 montre la base de données simplifiées. Les noms d'abstractions en majuscule dans ce schéma sont associés à l'entité éponyme. Chaque entité de type *FERME* contient une ou plusieurs entités *BASSIN*. Chaque *BASSIN* sera associé à un ou plusieurs *ELEVAGE* réalisés successivement au cours du temps. La durée d'un *ELEVAGE* est définie entre l'instant d'ensemencement et une *Vidange* dite finale. Chaque *ELEVAGE* sera caractérisé par les entités *VARIABLE DE PRODUCTION* et *VIDANGE*. Une *VIDANGE* est caractérisée par des *VARIABLE DE QUALITÉ*.

Quelques catégories liées à la *VARIABLE DE PRODUCTION* sont listées ci-dessous :

- ***Indicateur de performance*** : variable relevée au cours de l'élevage qui permet de fournir une indication temporelle sur le développement de la production (croissance, mortalité...)

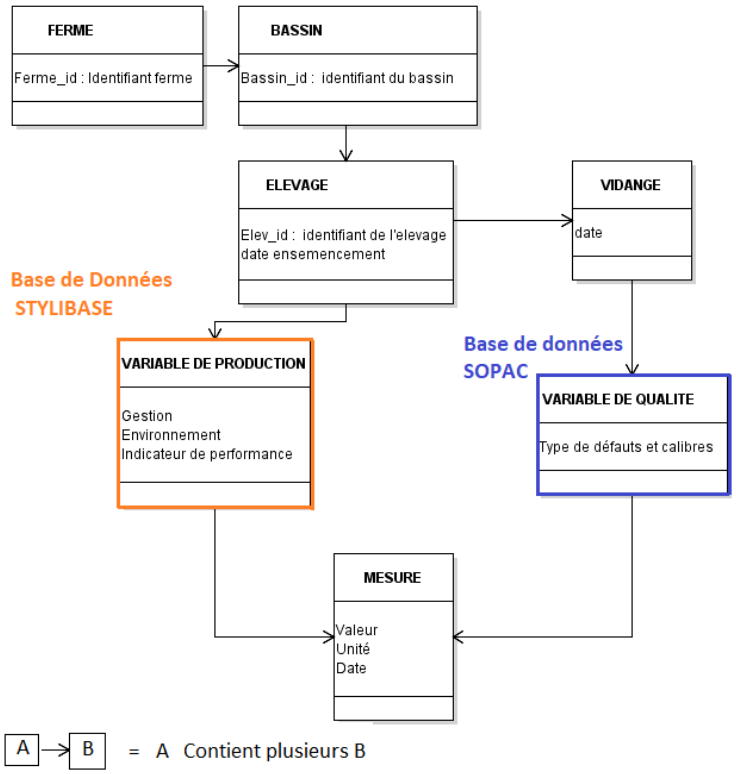

Fig. 4.8 Base relationelle simplifiée des données étudiées

- ***Variable de gestion*** : variable temporelle sur laquelle le fermier est en capacité d'agir directement sur sa valeur à un instant *t* défini et qui peut avoir une influence sur la qualité de la production.
- ***Variable d'environnement*** variable temporelle décrivant l'évolution de la qualité du milieu d'élevage.

Enfin, les entités *variable de production* et *variable de qualite* possèdent une série de *mesures*. Chaque *mesure* est associée à une date de prélèvement, une unité et un type de données parmi les types cités précédemment (catégoriel, quantitatif, empirique).

4.1.3 La quantité de mesures

La figure 4.9 présente le nombre de mesures d'élevage présentes par année dans *STYLIBASE* entre 2000 et 2016. Elles ont une fréquence horaire, journalière ou hebdomadaire. Pour un total d'environ 1.2 millions de mesures, le nombre de mesures varie progressivement de 2000 à environ 75000 mesures par année, entre 2000 et 2005. On remarque ensuite une quantité plus importante de mesures disponibles :

- de 2007 à 2009 avec une moyenne d'environ 90000 mesures par année.
- de 2011 à 2013 avec une moyenne d'environ 100000 mesures par année.

- en 2015 avec près de 100000 mesures.

Sur la période allant de 2005 à 2016, on remarque un nombre disponible de mesures qui décroît légèrement en 2006, 2010 et 2016 pour atteindre une moyenne de 65000 mesures enregistrées par an.

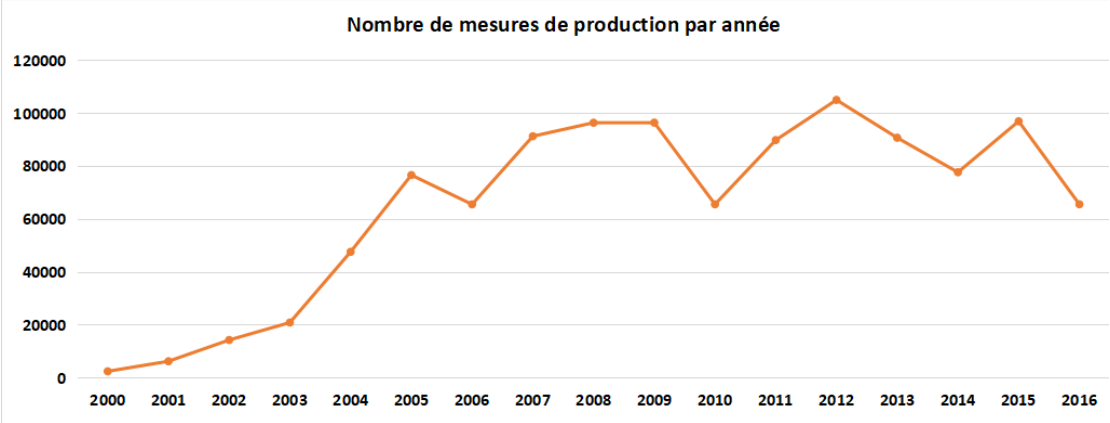

Fig. 4.9 Nombre de mesures dans STYLIBASE par année

La figure 4.10 présente la quantité moyenne de mesures par mois. En moyenne mensuelle, entre 2000 et 2016, les mois compris entre juin et octobre comptabilisent un nombre moins important de mesures enregistrées. Cette baisse s'explique par un arrêt de la production en période de saison fraîche due à une vibriose dénommée Syndrome 93. L'émergence de cette maladie en 1993 a conduit la filière à concentrer sa production sur la période estivale.

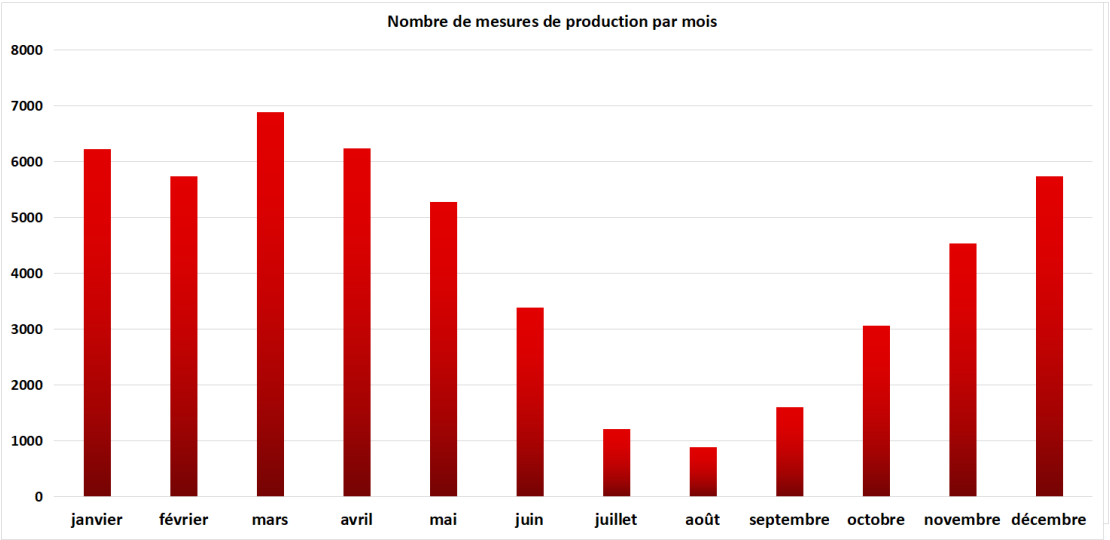

Fig. 4.10 Nombre de mesures dans STYLIBASE par mois

Pour 700 élevages enregistrés dans *STYLIBASE*, des statistiques descriptives ont été utilisées afin de donner une vision globale de l'évolution de la base de données par

variable entre 2000 et 2016. La précision de certaines de ces mesures, par heure, est fournie. Ces détails permettront au lecteur de se rendre compte de l'absence de norme standard de suivi des variables entre les fermes, et que leurs analyses à l'échelle de la filière est complexe.

A titre d'exemple, l'horaire matinale dédiée aux mesures de température, varie en fonction des fermes. Des séries de données représentant des valeurs journalières moyennes de ces variables peuvent être générées et utilisées, dans le cadre d'une analyse de séries temporelles mono-variées à l'échelle de l'ensemble des élevages et sur une échelle de temps commune. Néanmoins l'analyses de ces valeurs moyennes, doit être complétée dans l'absolu par une analyse des séries de données matinales, ou relevées en fin de journée afin d'évaluer l'impact de la variation de ces variables sur la performance d'élevage.

La figure 4.11 affiche l'évolution des moyennes mensuelles des variables de production recueillies durant les élevages réalisés entre les années 2000 et 2016.

Après une augmentation du nombre de mesures, entre 2001 et 2007 quelle que soit la variable, on note une stabilisation qui s'explique par un nombre de fermes qui reste constant.

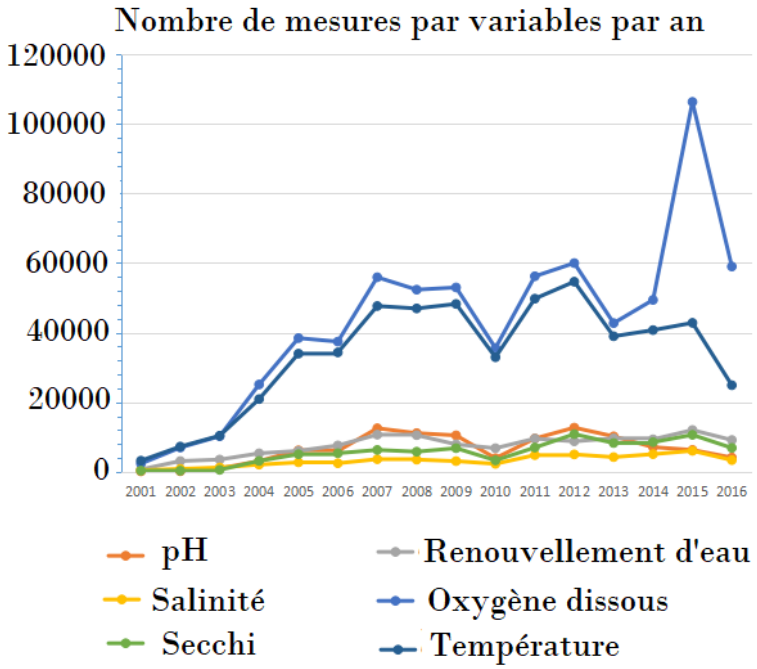

Fig. 4.11 Évolution du nombre de mesures par variable par année

On remarque deux périodes importantes pour le recueil de données pour l'ensemble des fermes de la filière :

- entre février et août d'une année n
- entre août de l'année n et février de l'année $n + 1$.

Ces deux périodes sont les deux périodes principales d'élevage. Dans la première période, l'ensemencement des élevages s'effectue durant la saison chaude, entre février et mars, durant laquelle, la température dans les bassins est généralement supérieure à 30°. Durant la seconde période les élevages débutent entre août et octobre et se termine au début de l'année suivante.

La fréquence d'acquisition de ces variables, durant la phase des élevages, varient considérablement entre chacune d'elles. On note aussi une variation très importante d'une ferme à l'autre.

4.2 Descriptif des variables environnementales

On peut distinguer plusieures classes de données en fonction de la fréquence des relevés et de la méthode d'acquisition. La première classe concerne la température et l'oxygène. Ces données sont généralement relevées par des ouvriers au levé et au couché du soleil. Le pH est généralement relevé par des techniciens au cours de la journée. La figure 4.12 affiche le nombre de mesures acquises et leur représentativité.

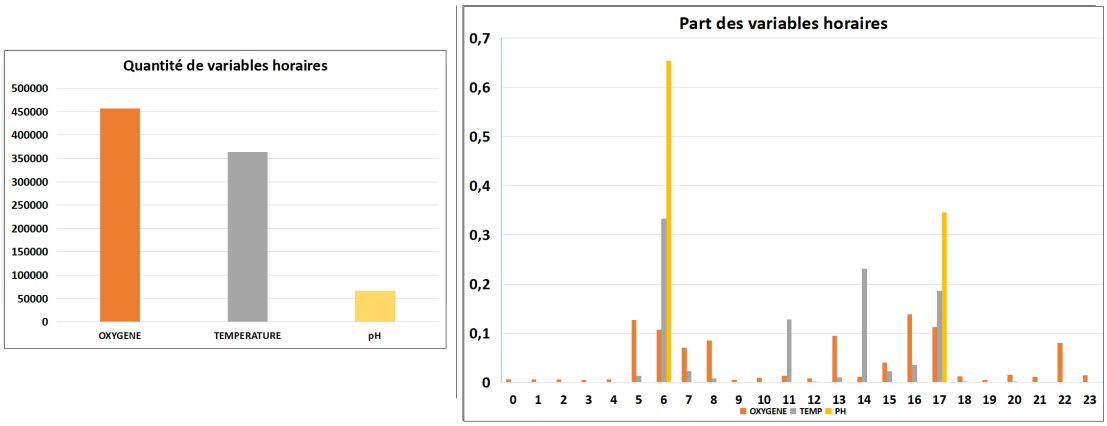

Fig. 4.12 Représentativité des variables horaire

Les mesures d'oxygène dissous sont relevées essentiellement en début et fin de journée lorsque les valeurs deviennent minimales et maximales. Il en est de même pour la température mais la répartition des valeurs semble plus étalée. Il est probable que la stratégie d'échantillonnage diffère d'une ferme à l'autre.

4.2.1 les principales variables environnementales mesurées

La figure 4.13 montre les statistiques pour 3 variables descriptives que sont la température, l'oxygène dissous et le pH. On notera une forte variation des valeurs quelle que soit la variable.

La figure 4.14, affiche le nombre de mesures par heure (à gauche) et par année (à droite) pour ces 3 variables. La figure 4.15, affiche le nombre de fermes enregistrant

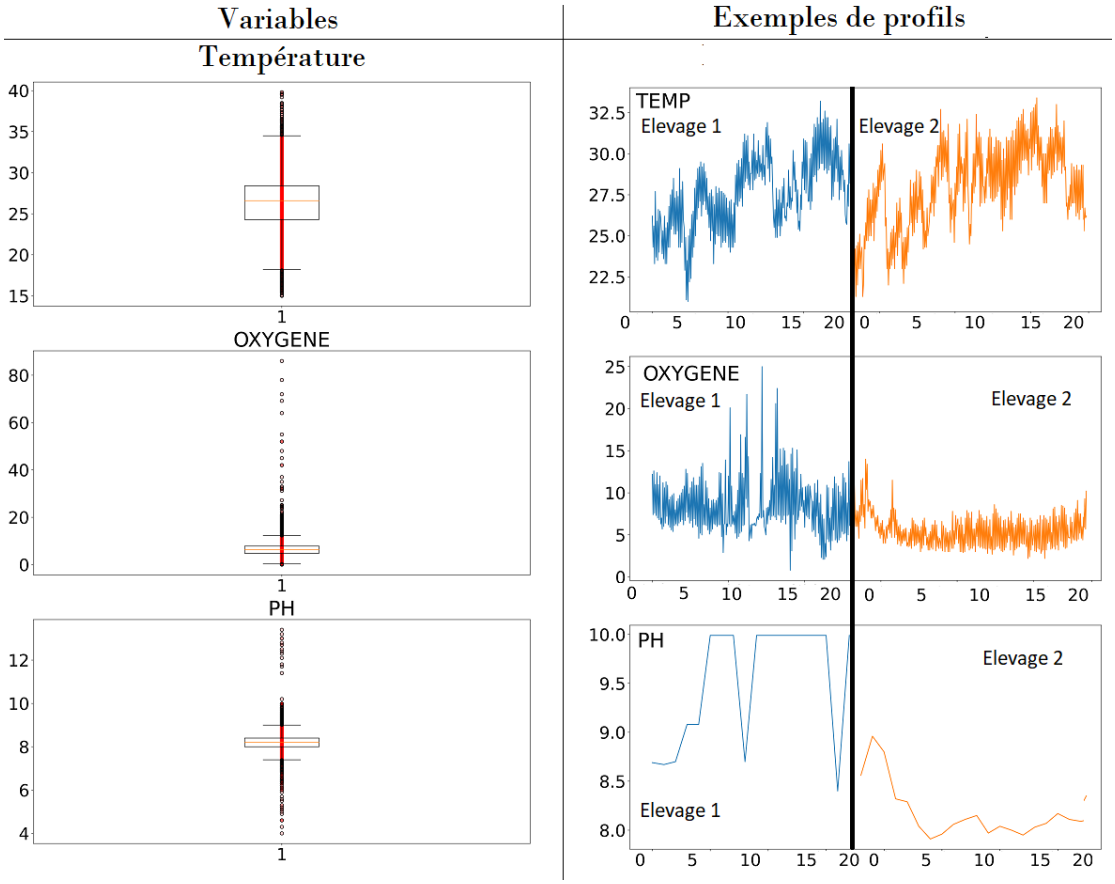

Fig. 4.13 Distribution des données des variables environnementales et de production ayant une fréquence d'acquisition horaire, prises sur l'ensemble des élevages. Exemple d'évolution sur les 20 premières semaines de deux élevages (à droite). La température est exprimée en °C, l'oxygène en mg/l et le pH en unité pH.

des valeurs en considérant 'heure (à gauche) ou l'année (à droite) d'acquisition. Ces variables ont une fréquence d'acquisition horaire.

La température: La température est davantage suivie à $6h$, $11h$, $14h$ et $17h$ avec des nombres de mesures respectifs de 94582, 39255, 65045 et 58415. Ces mesures ont principalement été relevées entre les années 2000 et 2016 (cf. figure 4.14).

Le nombre important de mesures, liées aux horaires les plus ciblées (cf. figure 4.14), peut, par comparaison avec la figure 4.15, être expliqué par un nombre de fermes plus important qui enregistrent des données à ces mêmes heures. La figure 4.15 montre que la majorité des fermes enregistrent (au moins) une mesure de température à $6h$ et à $14h$. Cette information est importante dans le cas où des normes d'élevage doivent être déterminées. Les séries temporelles provenant de toutes les fermes, doivent si possible être analysées selon des échelles et des fréquences d'acquisition comparables. Néanmoins nous verrons en fin de chapitre, qu'il existe un fort déséquilibre au niveau de la représentativité des données de ces fermes.

Le nombre d'enregistrements de mesures de température d'eau, est plus important

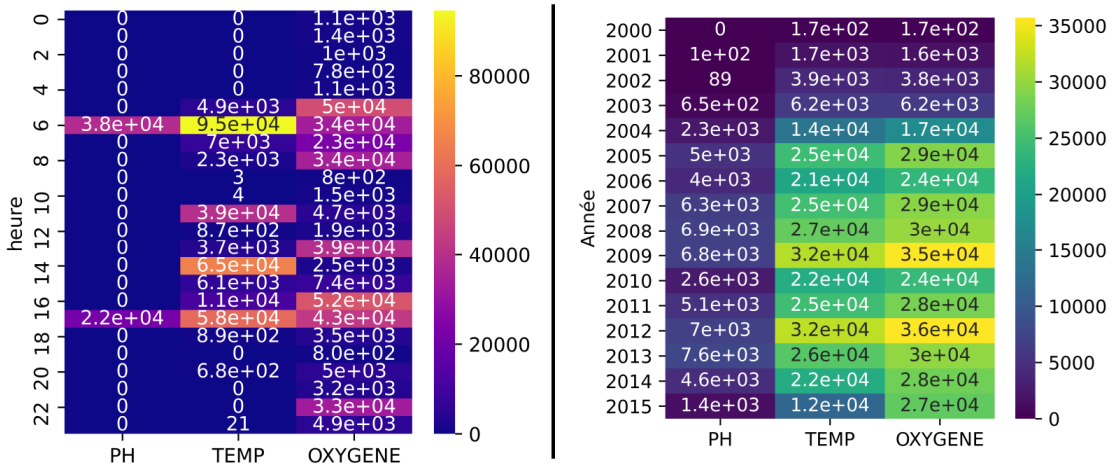

Fig. 4.14 Nombre de mesures relevées en fonction des heures d'acquisition, et des années, pour les variables environnementales.

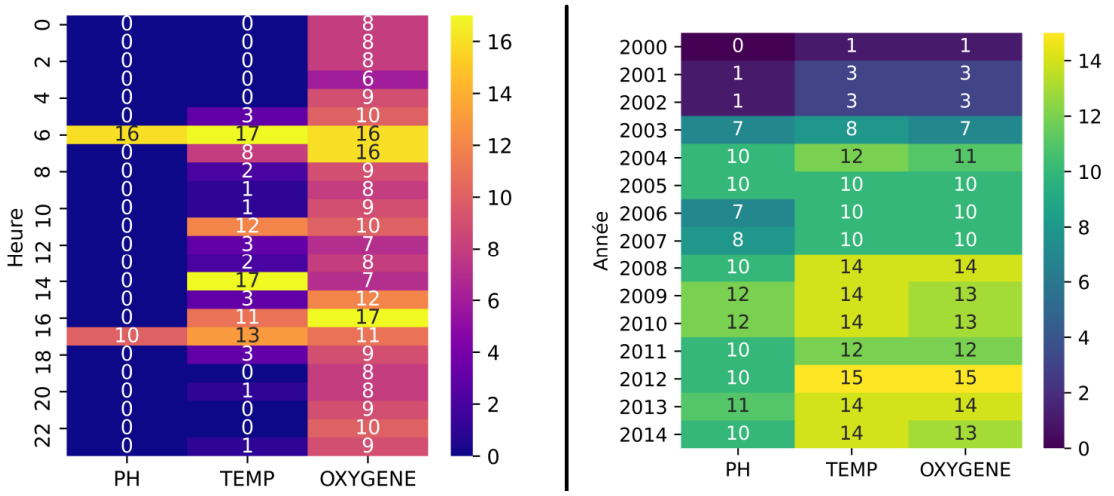

Fig. 4.15 Nombre de fermes, enregistrant des données par année pour les variables pH, Température et oxygène.

entre les années 2004 et 2014 (cf. figure 4.14). Ce nombre est lié à celui des fermes enregistrant des données à cette même période. Il est judicieux de considérer cette fenêtre temporelle pour une analyse de la température à l'échelle de la filière aquacole Calédonienne, afin de déterminer des normes représentatives de l'évolution de la qualité d'eau des bassins de la filière.

La figure 4.16 présente l'évolution de la température moyenne enregistrée à l'échelle d'une semaine. L'abscisse exprime le nombre cumulé de semaines au cours d'une même année. Sur cette même figure, pour chacune des heures les plus suivies ($6h$, $14h$, $11h$, $16h$ et $17h$) les valeurs moyennes, minimales, maximales (de la température de l'eau des bassins), ont été calculées. Par exemple pour obtenir Les valeurs moyennes pour l'horaire 14, les valeurs relevées à 14 heures durant la première semaine ont été

considérées dans le calcul. Cette figure montre que ces évolutions sont représentatives du climat en Nouvelle-Calédonie. Le fermier ne peut agir directement sur cette variable.

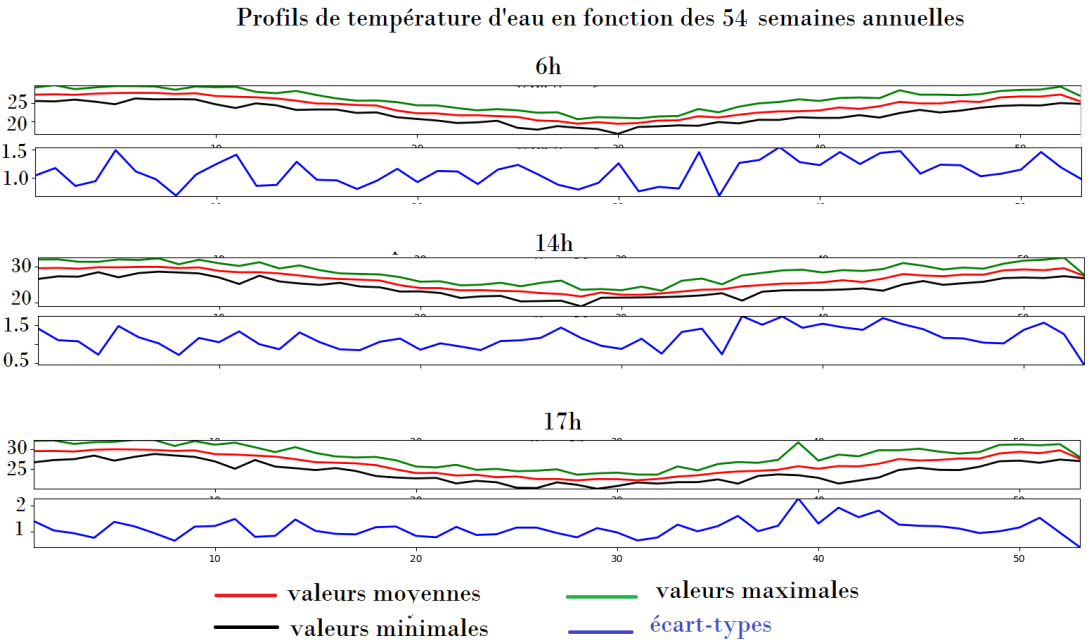

Fig. 4.16 Évolution de la température moyenne par semaine annuelle

L'oxygène dissous : 400 élevages suivis entre 2000 et 2016, montrent des mesures d'oxygène dissous dans *STYLIBASE*. Les valeurs sont enregistrées généralement entre $5h$ et $8h$, à $13h$, entre $16h$ et $17h$ et à $22h$.

Le pH : 290 élevages suivis entre 2000 et 2016, montrent des mesures du pH, dans *STYLIBASE*. Le suivi de cette variable n'est pas réalisé dans toutes les fermes.

Le pH est plus particulièrement suivi:

- à $6h$ avec 38326 mesures
- à $17h$ avec 22055 mesures

Le nombre de mesures de pH à $06h$ heure la plus suivie peut être expliqué, comme pour la température, par un nombre de fermes plus important l'enregistrant.

Comme pour les autres variables dès 2004 les données sont davantage renseignées dans *STYLIBASE*, et nos analyses seront effectuées sur les données acquises à partir de cette année.

La quantité de mesures n'est pas associée au nombre de fermes relevant cette donnée, comme pour la température.

4.2.2 Les variables environnementales secondaires

Les trois variables environnementales avec un nombre de mesures moins importants sont le disque de secchi, la salinité de l'eau, et la fluorescence. Pour ces trois vari-

ables, la figure 4.17 affiche, par année, le nombre de fermes enregistrant des données (à gauche), et le nombre de mesures (à droite).

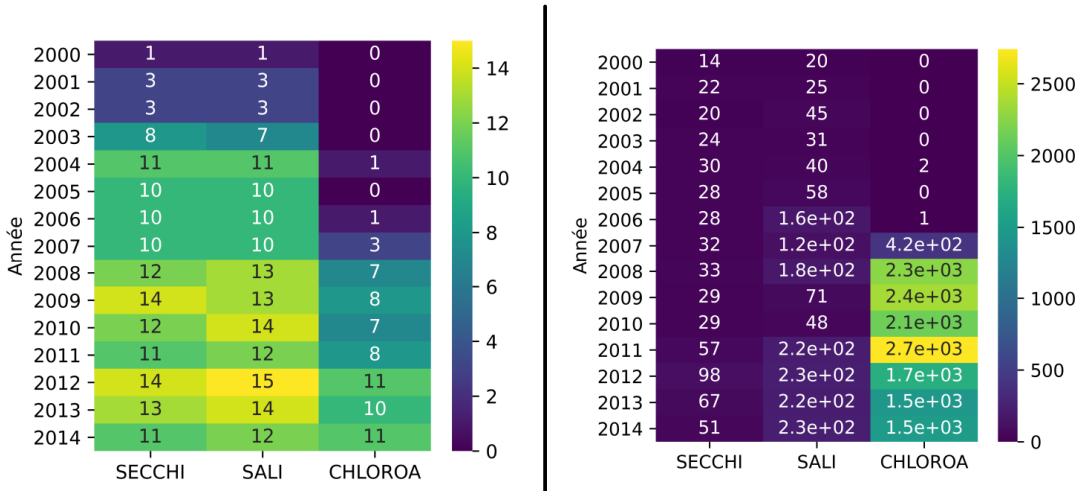

Fig. 4.17 Nombre de mesures enregistrées et des fermes relevant les données en fonction de l'année calendaire

Le secchi : 373 élevages suivis entre 2000 et 2016, montrent des mesures du disque de secchi, dans *STYLIBASE*.

La salinité : 295 élevages suivis entre 2000 et 2016, ont enregistrés des mesures de salinité de l'eau, dans *STYLIBASE*.

La fluorescence : 239 élevages suivis entre 2000 et 2016, ont enregistrés des mesures de fluorescence, dans *STYLIBASE*.

La figure 4.18 affiche le nombre de mesures pour ces variables. Contrairement aux variables environnementales précédentes (température, oxygène dissous et *pH*), ces variables sont représentées par au maximum une seule valeur par jour. L'heure du relevé peut être différente dans la journée.

La figure 4.19 montre les intervalles de valeurs des variables journalières, obtenus sur l'ensemble des élevages qui les enregistrent (box-plot de gauche). A droite de la figure, il y a des exemples d'évolution de ces variables sur les 20 premières semaines de deux élevages.

4.3 Descriptif des variables de gestion

La figure 4.20 présentent des informations descriptives pour les deux principales variables de gestion : l'aliment et le renouvellement.

Pour ces deux variables, une évolution croissante est visible jusqu'en milieu d'élevage. Ensuite les taux de renouvellement de l'eau et d'alimentation décroissent suite aux premières pêches qui commencent autour de J120. Cette gestion est similaire à toutes

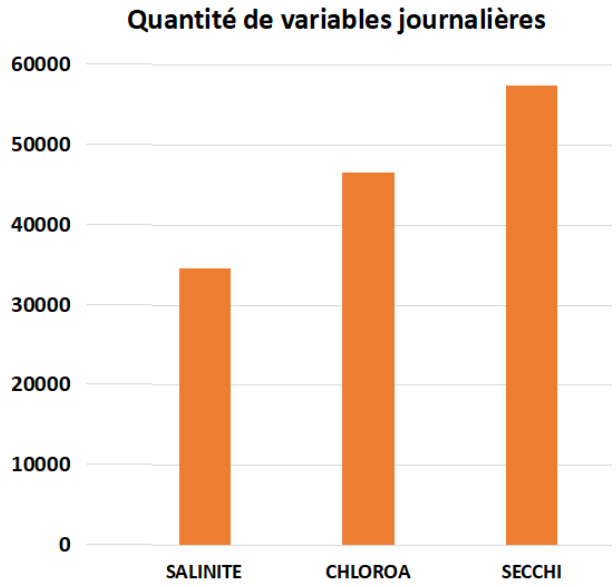

Fig. 4.18 Représentativité des variables environnementales les moins suivies

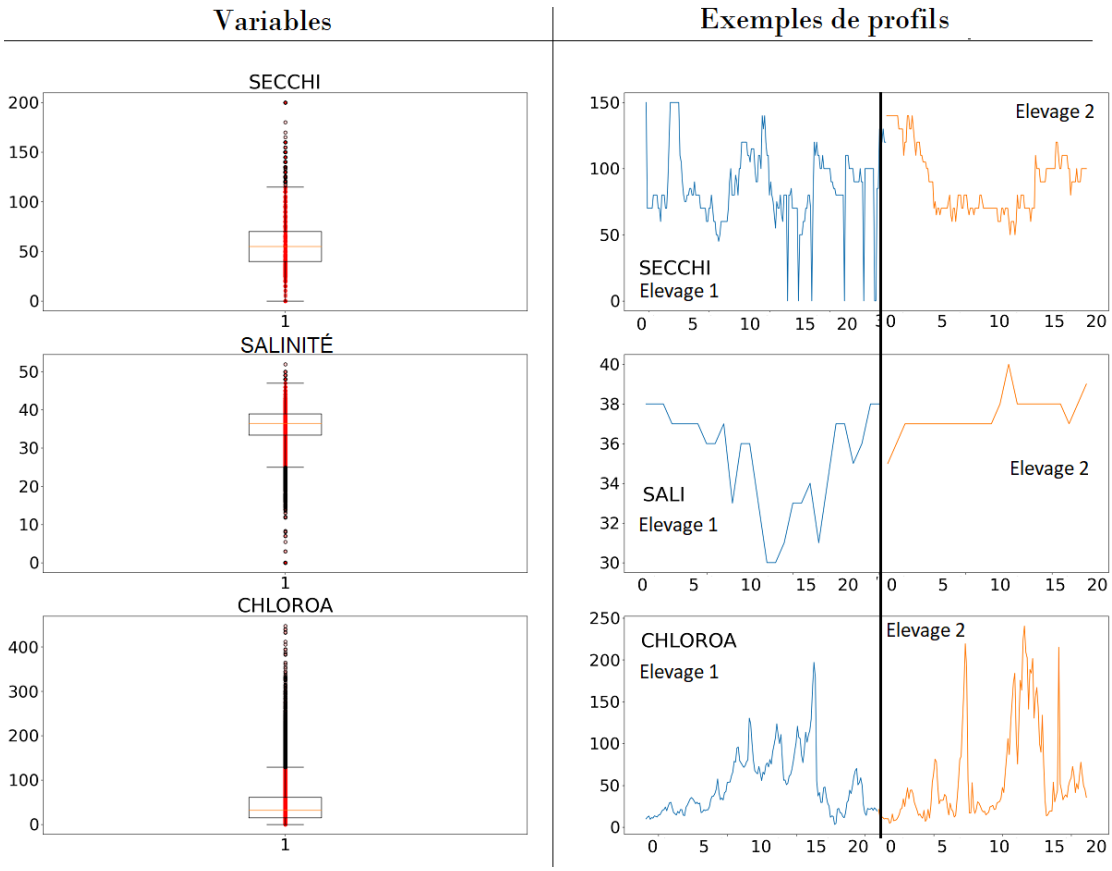

Fig. 4.19 Distribution des variables environnementales et de production mesurées avec une fréquence d'acquisition journalière. Exemple d'évolution sur les 20 premières semaines d'élevage pour deux élevages (à droite). Le secci est exprimé en cm, la salinité en PSU et la fluorescence en µg/L.

les fermes. Néanmoins le taux de renouvellement et de nourrissage divergent en fonction de la taille du bassin et parfois de l'appréciation de l'aquaculteur.

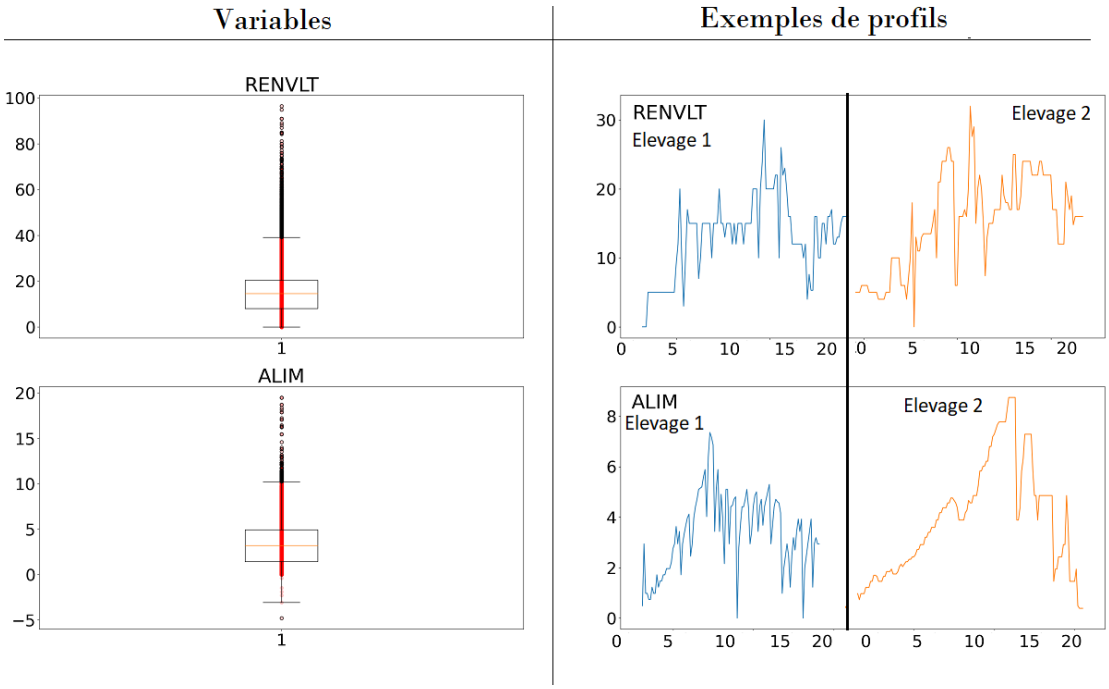

Fig. 4.20 Distribution des données des variables temporelles de gestion, prises sur l'ensemble des élevages. Exemple d'évolution sur les 20 premières semaines d'élevage pour deux élevages (à droite). . Le renouvellement est exprimé en % et l'aliment en g/m²/jour.

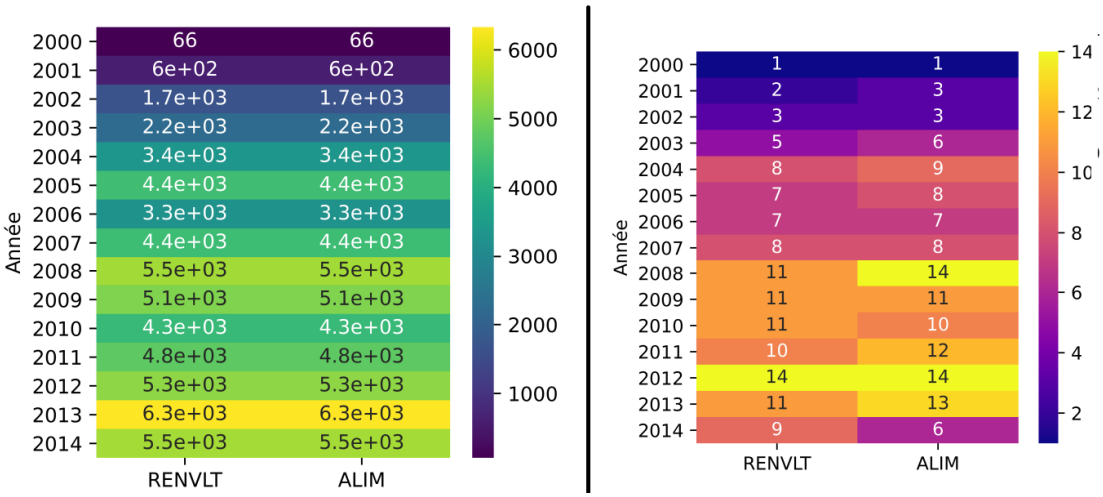

Fig. 4.21 Nombre de mesures enregistrées, et des fermes relevant les données en fonction de l'année d'élevage, par variable de gestion

Le renouvellement de l'eau : 490 élevages suivis entre 2000 et 2016 montrent des valeurs pour le renouvellement d'eau.

4.4 Les données de performance d'élevage

4.4.1 Le poids moyen

Durant l'élevage, un échantillon d'animaux est prélevé de manière hebdomadaire pour calculer un poids moyen, en plusieurs points considérés par l'éleveur comme échantillons représentatifs. Comme énoncé dans les chapitres précédents, cette variable zootechnique est un facteur prédominant pour l'évaluation de la performance de l'élevage.

La figure 4.22 montre les intervalles de valeurs des données de croissances, prises sur l'ensemble des élevages (boxplot de gauche). Des courbes représentatives de croissance sont affichées à droite de la figure 4.22. On remarque une convergence des courbes de croissance vers un poids (moyen) final. Ces courbes suivent des tendances que l'on peut modéliser. Nous verrons dans le chapitre suivant, dans lequel un modèle de croissance non linéaire est appliqué, comment des paramètres zootechniques de ces courbes seront extraits, afin de proposer des normes de croissance pour cette filière aquacole de Nouvelle-Calédonie.

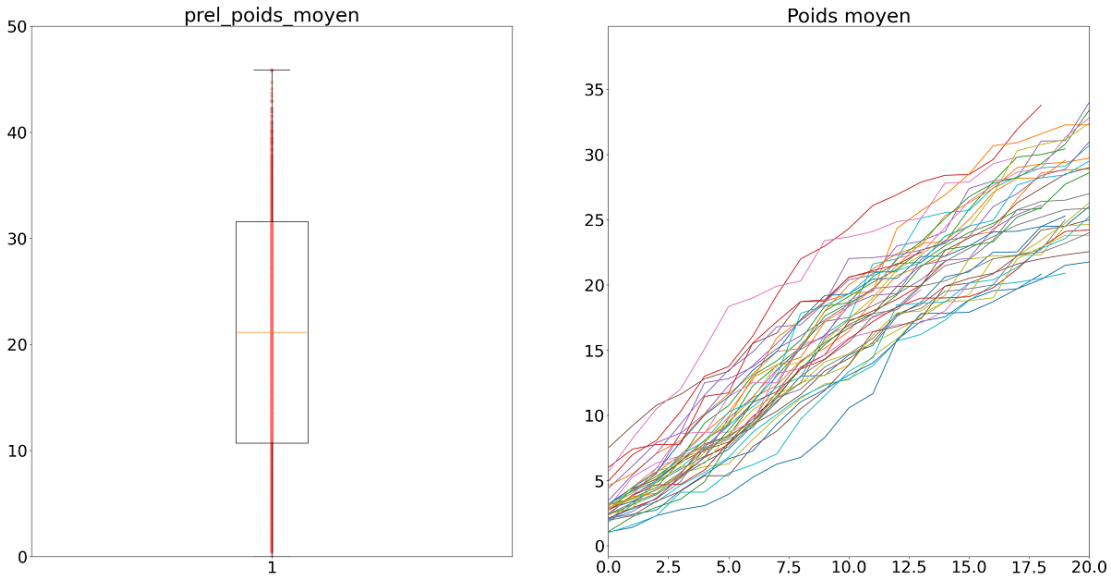

Fig. 4.22 Distribution des données poids prises sur l'ensemble des élevages. Exemple d'évolution de la croissance des animaux sur les 20 premières semaines pour deux élevages (à droite).

4.4.2 La survie

La survie est un facteur prépondérant pour la rentabilité des filières. Elle dépend d'un ensemble de facteurs, dont les plus importants sont liés à la qualité du milieu (eau et sédiment) et des paramètres de gestion (densité initiale, taux de renouvellement, taux d'alimentation....).

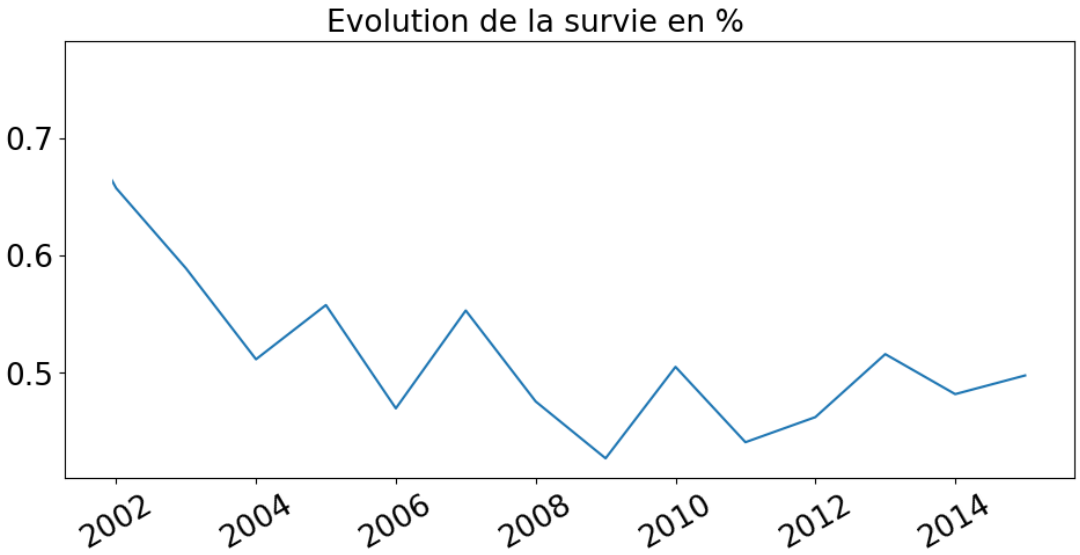

Fig. 4.23 Évolution du taux moyen de la survie des productions de crevette entre 2000 et 2015

4.5 Données de qualité du produit relevées par la SOPAC

Les données de qualité font principalement références à des défauts visibles, et sont relevées sur la crevette à chaque pêche. Pour rappel, il y a en moyenne 3 pêches par élevage qui sont transportées vers la société SOPAC, dont 2 pêches intermédiaires et 1 pêche dite finale. La dernière pêche appelée pêche finale conduit à la vidange totale du bassin.

4.5.1 Processus d'évaluation de la qualité de la production

La venue des crevettes, depuis leur ferme de grossissement, se fait dans des contenants de 4 tonnes (Cf. image de gauche de la figure 4.24). Chaque contenant est lié à la production d'un élevage particulier, et qui est associé à l'identifiant de la base de données *STYLIBASE*. Un second identifiant enregistré par la *SOPAC* est conservé par le Groupement des Fermes Aquacoles (GFA). Ces identifiants permettent le traçage des produits de la pêche.

Le croisement de deux identifiants pour relier les deux bases a fait l'objet d'un travail fastidieux au cours de cette thèse. Ce travail a été effectué pour retrouver les données de production, à partir d'informations incomplètes.Le croisement a été validé avec certains membres de la SOPAC, de l'équipe LEAD de l'IFREMER en charge du développement de Stylog et du GFA. Il a permis de relier la qualité du produit pêché à son élevage sans identifier clairement les fermes et les bassins afin de respecter la clause de confidentialité imposée par le GFA (Cf section 4.1.0.1).

Pour déterminer la qualité du produit, plusieurs contrôles sont effectués à l'arrivée

Fig. 4.24 Caisse de transport des crevettes d'une ferme à l'usine

en usine. Plusieurs échantillons (Cf. figure 4.25) sont prélevés à plusieurs endroits dans chaque conteneur et à chaque endroit du conteneur (au dessus, au fond et sur les cotés).

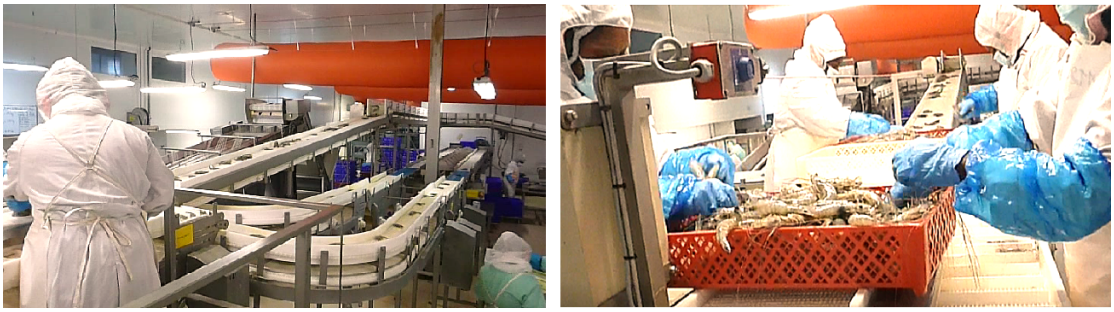

Fig. 4.25 Tri des crevettes selon les défauts visibles

A partir de ces échantillons, une analyse en laboratoire est réalisée afin de rechercher la présence de défauts, et de maladies, pour estimer la qualité de la production. Environ $5kg$ par box sont inspectés par lot (production d'un même bassin). Une extrapolation de cette estimation permet de compléter un document officiel sur la qualité du produit par lot et par ferme.

4.5.2 Les différents défauts relevés

Le tableau 4.5 montre les types de défauts relevés par la SOPAC, dont certains sont aussi présentés sur la figure 4.26.

Les défauts relevés peuvent apparaître sur la tête et la queue de l'animal. Une coloration orange à rouge est l'un des défauts les plus communs. Les défauts associés à la présence de cicatrices, de points noirs, d'une déformation de l'animal sont relevés. Le classement des défauts est fondé uniquement sur l'expertise des responsables qualités de la société *SOPAC*.

L'historique de ces informations récoltées depuis l'année 2000, a été transmis par

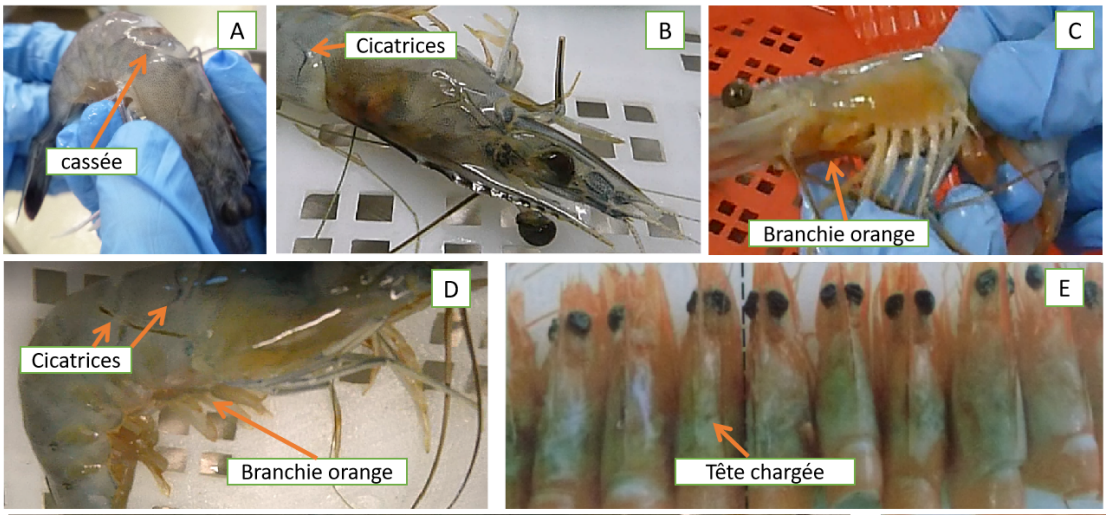

Fig. 4.26 Quelques défauts visibles sur la crevette, avant ou après cuisson

Abréviation SOPAC	Description
Défauts sur la tête	
Tête_rouge	Coloration rougeâtre sur la tête
Tête chargée	Liquide verdâtre visible après cuisson
Tet_Eclat_Crues	tête éclatée avant cuisson
Tet_Eclat_Cuites	tête éclatée après cuisson
Défauts sur le corps	
Def_Branch	Défaut relevé sur les branchies
Def_carapace	Défaut relevé sur la carapace
Branchies_foncees	Coloration visible sur les branchies
Pat_Rouges	Coloration rougeâtre sur les pattes
Pat_Vert	Coloration verdâtre sur les pattes
Autres types de défauts	
Deformees	Crevette déformée
Cica	Grosses et petites cicatrices visibles
Rostre_casse	Rostre cassée

Table 4.5 Défauts relevées par la société *SOPAC*

la SOPAC, dans un fichier au format textitExcel. Ce fichier contient les informations recueillies directement par l'usine. Il concerne la qualité de plus de 5900 pêches. Un extrait de ce fichier est montré figure 4.27. Les informations incluent également :

- l'heure et la date d'arrivée et de réception de la marchandise à l'usine, *SOPAC*
- l'heure et la date de début et de fin de traitement et de conditionnement des productions,
- l'identifiant SOPAC de la production
- Le pourcentage de chaque défaut relevé visuellement (Cf. tableau **??**),

- Le pourcentage de crevettes dites 'Bas de gamme',
- Le pourcentage de crevettes dites 'bonnes' qui ne montrent aucun défauts ou dont la présence de défauts est inférieure à un seuil acceptable,
- L'état de l'exosquelette et qui est fonction du stade de mue: molle, carto (plus de solidité).

A	B	C	D	E	F	G	H
Annee_recep	Code_Batch	Mois_recep_	Date_Recep du camion	Heure Recept du camion à l'usine	Heure_Echant	Heure_Deb_Production	Heure_Fin_Produc
2000	A010178	06	20000626	13:30:00	15:00:00	18:00:00	22:45:00
2001	A011143	05	20010523	16:30:00	13:55:00	20:15:00	22:15:00
2001	A011156	06	20010605	08:15:00	09:30:00	10:20:00	13:50:00
2001	A011180	06	20010629	06:45:00	07:00:00	07:45:00	13:00:00
2001	A011207	07	20010726	05:30:00	05:45:00	06:20:00	09:55:00
2001	A011219	08	20010807	05:30:00	05:50:00	06:15:00	08:05:00
2002	A012352	12	20021218	07:00:00	07:15:00	10:20:00	12:15:00
2003	A013003	01	20030103	07:00:00	07:30:00	09:50:00	13:20:00
2003	A013027	01	20030127	07:00:00	07:50:00	10:40:00	13:05:00

N	O	P	Q	R	S	T	U	V	W	X
Bonnes	Premium	Basdegamme	Molles	Carto	Melanosees	Def_Branch BDG	Cica_Gros BDG	Tet_Rouge_Deb	Tet_Eclat_Crues BDG	Tête chargée
40,72	47,36	11,92	4,72	30,24	0	0	3,84	17	2,6	0
6,6	73,2	20,2	0	12,56	0	0	10,2	12,88	0,4	0
17,32	68,56	14,12	1,4	45,52	0	0	2,6	42,52	1,32	0
62,5	14,76	22,72	0	3,6	0	0	0	1,6	2,08	0
50,16	27,72	22,12	4,64	3,56	0	0	1,6	4,92	0	0
11,8	58,8	29,4	11,32	23,24	0	0	0	5,68	0,64	0
24,72	34,84	40,44	1,08	5,48	0	0	20,4	27,44	0,92	0
4,28	77,4	18,3	0	14,64	0	0	2,8	33,64	0	0
43,16	13,8	43,04	8,72	14,84	0	0	19,48	0,84	1,2	0

Fig. 4.27 Extrait des données récoltées sur de la qualité du produit à chaque pêche

La figure 4.28 affiche, par type de défauts relevées dans le tableau 4.5, la moyenne annuelle des estimations de leurs apparitions par l'ensemble des élevages. On constate que la majorité des défauts a été recensé dès 2000 à l'exception :

- des défauts : "*Tête chargée*" et "*Tet_Eclat_Cuites*" recensés à partir de 2010,
- des défauts "*Rostre_casse BDG*" et "*Branchies_foncees_BDG*" recensés à partir de 2014.

Le recensement de ces défauts fait suite à une apparition significative relevée dans un ou plusieurs élevage, conduisant à leurs enregistrements dans la base de données *SOPAC*. On peut citer en particulier l'apparition de branchies orange dans la fiche qualité à partir de 2014. Ce défaut est lié à une acidification des sédiments qui conduit à une augmentation de la concentration en fer sous sa forme réduite dans le milieu. La coloration des branchies s'explique par une précipitation de ce fer au niveau des lamelles branchiales [106].

On remarque, pour ces 5900 pêches, que les défauts observés sur la tête de l'animal, (recensés à partir de 2010), sont plus fréquents que ceux sur le reste du corps dans les données.

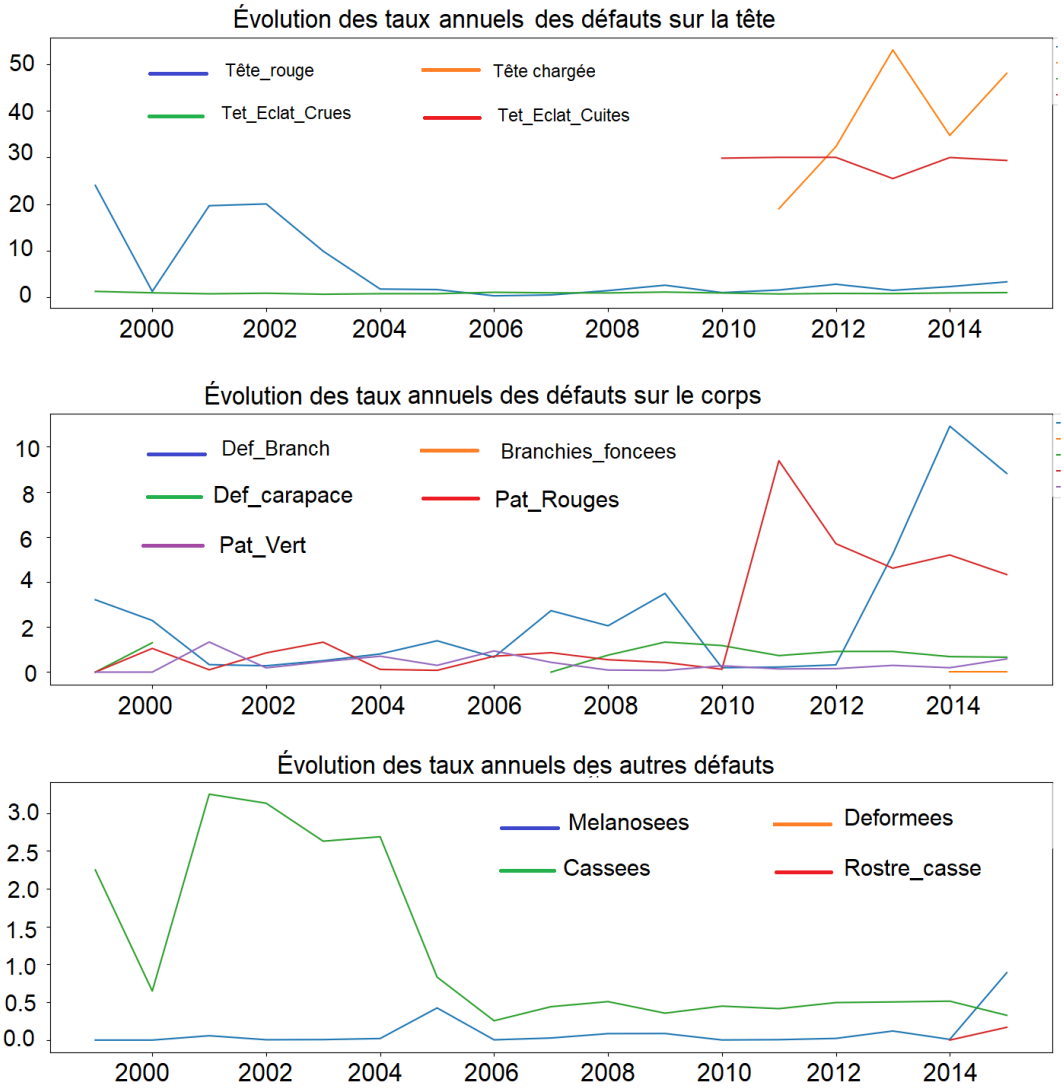

Fig. 4.28 Évolutions des taux d'apparition des défauts dans les productions

4.5.3 Les calibres

La *SOPAC* identifie différents calibres de crevettes. Le calibre $16/20$, par exemple, signifie qu'il faut 16 à 20 crevettes pour obtenir un poids d'un kg. Le prix payé aux fermiers dépend fortement du poids moyen des animaux et donc des calibres. A l'arrivée à l'usine, chaque lot est séparé en fonction des calibres des crevettes le composant. Les calibres relevés sont décrits dans le tableau 4.6. A droite du tableau, une photographie montre des crevettes de calibre 51/60.

La description fournie dans les sections précédentes, des données standards et de qualité de production, a mis en évidence la complexité de ces données. En effet ce sont principalement des données multi-variées et des séries temporelles multi-échelles (multi-variées). Certaines d'entre elles sont manquantes et imprécises.

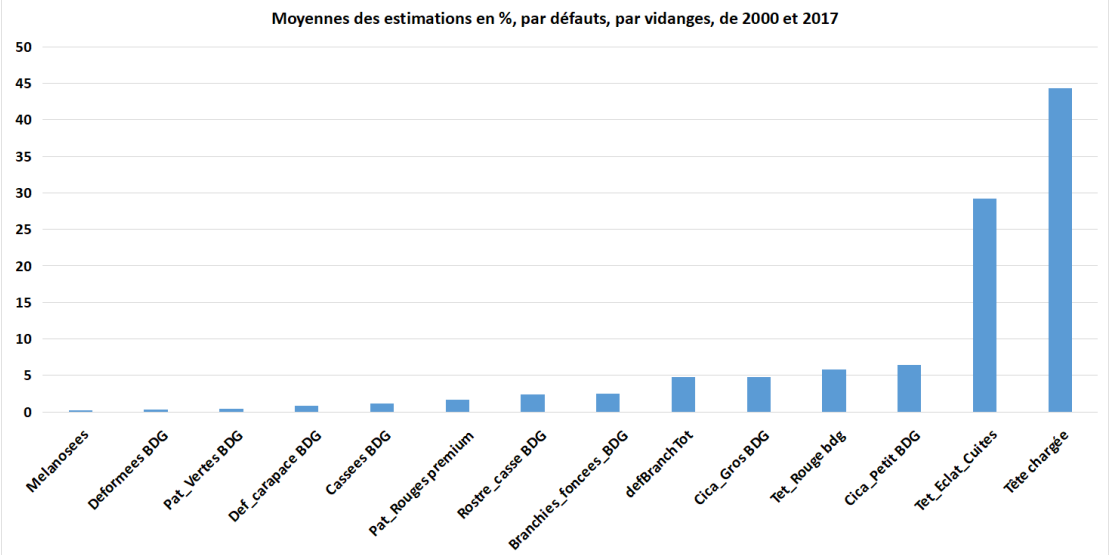

Fig. 4.29 Estimation moyenne par défauts

Fig. 4.30 Identification automatique des calibres

Calibres
16/20
21/25
26/30
31/40
41/50
51/60
61/80
+81

Table 4.6 A gauche la tableau des calibres, à droite un exemple de crevettes de calibre 51/60

4.6 imprécision et déséquilibre dans les données

La figure 4.31 présente une indication sur la représentativité des fermes, selon :

- le nombre total de mesures renseignées dans STYLIBASE (en orange)
- et le nombre total de mesures de qualité fournies par la SOPAC (en bleu).

On remarque une sur-représentativité des données de qualité et des données d'élevages sur certaines fermes. La *FERME11* a enregistrée par exemple plus de 45% du nombre total des données de production et 30% du nombre total des données de qualité).

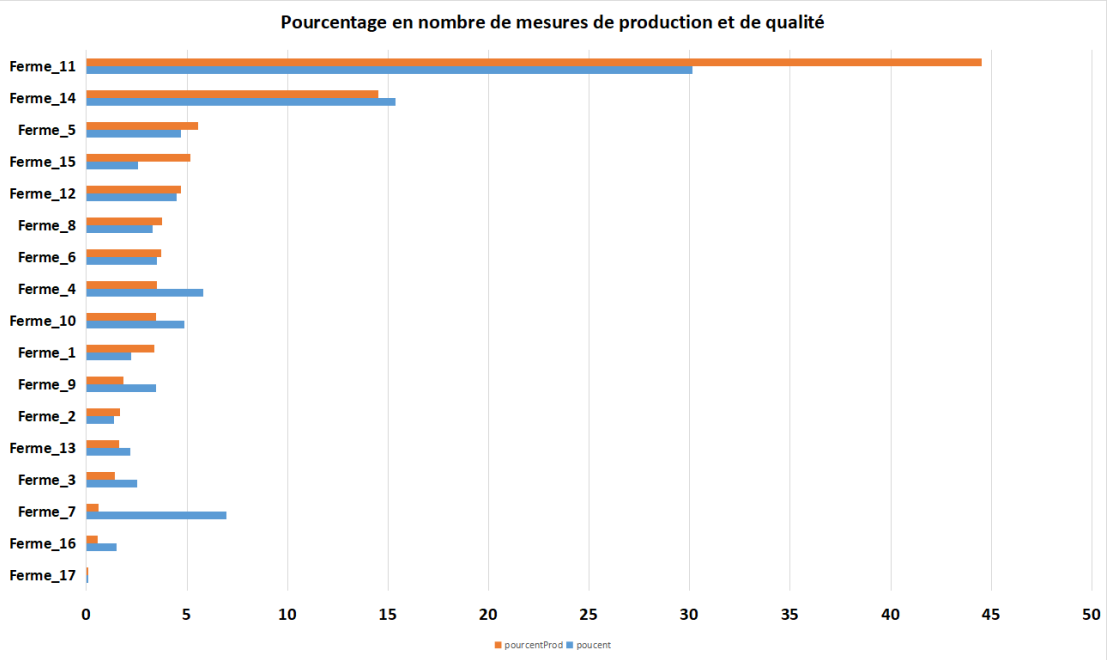

Fig. 4.31 Représentativité par ferme en pourcentage de données disponbles

70 bassins sont répertoriés dans *STYLIBASE*. La figure 4.32 affiche le nombre de données recueillis par bassin entre 2000 et 2016. Les points noirs dans la figure représente le nombre d'élevages par bassin, selon l'axe des ordonnées secondaire (à droite).

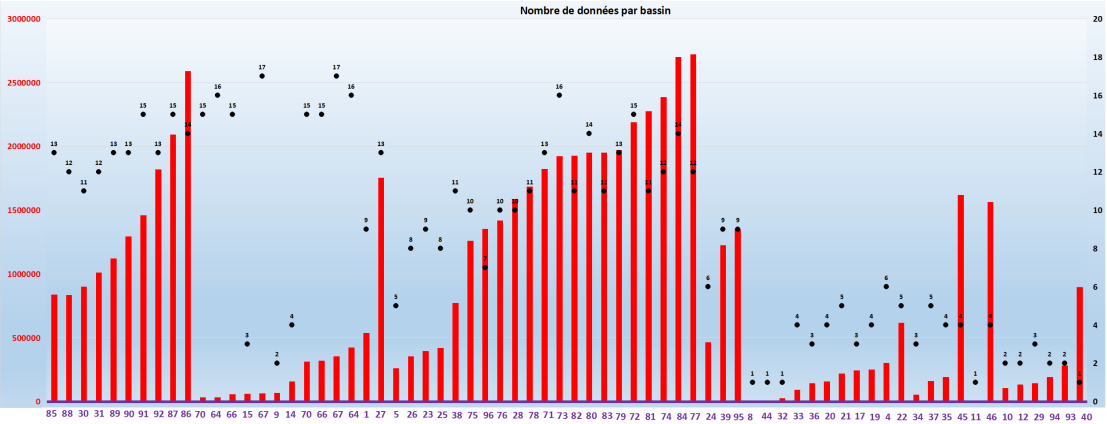

Fig. 4.32 Nombre de mesures par bassin

La figure 4.33 affiche, pour chaque ferme, la quantité de mesures par variables environnementales (présentées dans le tableau 4.4). Les mesures, pour l'oxygène dissous et la température, sont les plus fréquentes.

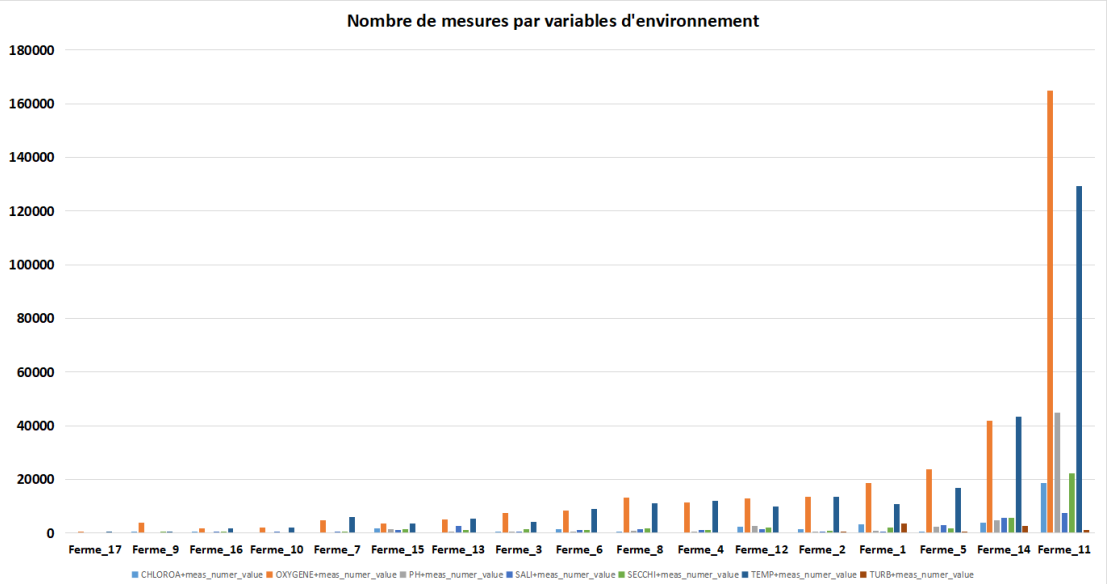

Fig. 4.33 Nombre de mesures par variable d'environnement par ferme

La figure 4.34 montre, pour chaque ferme, la quantité de mesures par variables de gestion (présentées dans le tableau 4.4).

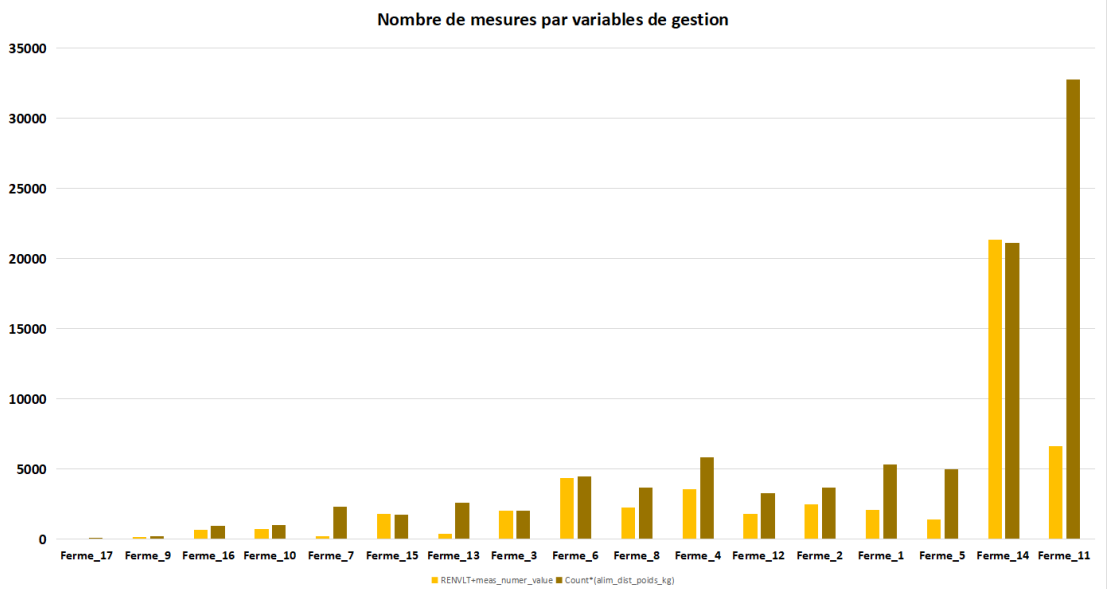

Fig. 4.34 Nombre de mesures par variable de gestion par ferme

Le début de la collecte des données pendant un élevage est très variable d'une ferme à l'autre. Certaines fermes choisissent d'enregistrer des valeurs, au remplissage du

bassin, d'autres après 15 à 20 jours après l'ensemencement. La figure 4.35 montre les valeurs des séries temporelles de température et d'oxygène dissous, enregistrées durant différents élevages.

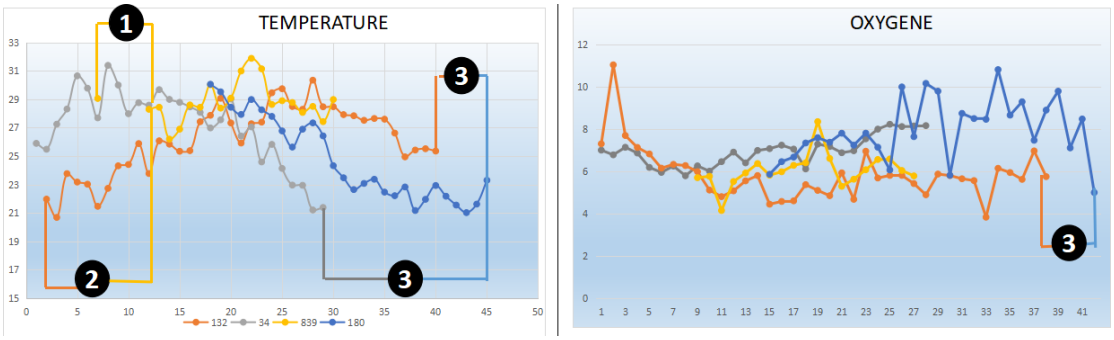

Fig. 4.35 Des séries temporelles déphasées

Les élevages recensés dans *STYLIBASE* montrent des phases de grossissement avec des durées hétérogènes. Néanmoins la durée minimale est d'environ 4 mois.

4.7 Conclusion

Nous remarquons que la collecte des données est fortement variable en fonction des fermes. Cette variabilité concerne aussi bien le nombre de variables collectées que la fréquence des relevés.

Pour les variables temporelles, la stratégie pour l'analyse des données sera, sachant que les fermes commencent la collecte des données à des moments différents, de déterminer un point initial commun pour comparer l'évolution des séries temporelles. Cette variabilité concerne aussi bien le nombre de variables collectées que la fréquence des relevés. Sachant que la fréquence d'acquisition des variables n'est pas la même en fonction des fermes, la comparaison entre fermes s'avère complexe. Il est donc nécessaire d'utiliser des méthodes qui soient suffisamment robuste pour être en mesure d'analyse le jeu de données dans son ensemble sans devoir supprimer un nombre trop important d'élevages. La classification supervisée ou non supervisée de séries temporelles multivariées multi-échelles et déphasées c'est à dire dont les données temporelles ne sont pas définies sur une fenêtre temporelle et une résolution temporelle communes, pour une même variable, et entre les différentes variables, n'a pas encore été proposée dans la littérature.

Nous proposerons ainsi dans le chapitre suivant, de créer de nouveaux descripteurs statiques et temporels pour permettre d'assurer des classifications supervisées, et non supervisées performantes, et adaptées à ces données complexes.

Chapitre 5

Stratégie d'analyse des données de la filière crevetticole Calédonienne

5.1 Contribution méthodologique et algorithmique pour l'analyse de données de filières aquacoles

La réussite d'un élevage aquacole dépend de nombreux facteurs, d'origine zootechnique ou économique. Les acteurs impliqués dans ce processus complexe de production doivent identifier les conditions favorables à l'optimisation du rendement et de la qualité du produit à l'échelle de la structure d'élevage, des fermes et/ou de la filière. La survie et la croissance des animaux restent par exemple des éléments incontournables qui vont dépendre de nombreuses variables. Ces filières produisent ainsi de nombreuses données, généralement sous-exploitées car complexes (hétérogènes, temporelles, spatiales, etc.) et issues de différentes sources (producteur, provendiers, sociétés de commercialisation...).

Dans ce chapitre, nous décrirons la démarche que nous avons développée pour analyser le jeu de données généré par de multiples acteurs et établi sur des élevages de la crevette tropicale Litopenaeus Stylirostris réalisés en Nouvelle-Calédonie entre 2000 et 2015. Le but de notre démarche est de croiser les données de production et de commercialisation afin d'identifier les meilleures pratiques zootechniques et les meilleures conditions environnementales possibles pour optimiser les rendements tout en générant un produit de qualité. Pour ce faire, nous proposons des scénarios méthodologiques de science de données (classification non supervisée et supervisée) sur les données produites par la filière pour d'abord (i) identifier les tendances ou les groupes de tendances des pratiques fermières les plus optimales, identifier les causes de mortalité et ensuite essayer de construire des modèles de prédiction. Vu la complexité des données, ces modèles seront appliqués à de nouveaux paramètres construits, par exemple à partir d'un modèle de croissance des animaux. Ces paramètres seront établis sur des données de poids mesurés par les éleveurs. Nous présenterons les résultats interprétés par les experts des données et nous discuterons de la suite de l'étude. Les méthodes *X-meansTS* et *X-meansMMTS* ont servi à analyser les données de qualité du milieu. Les clus-

ters obtenus ont été décrits par les données de qualité d'élevage, et par les paramètres qui ont construits à partir des données de croissance. Une Nouvelle méthodologie proposée, assurera l'extraction de nouvelles connaissances à partir de l'ensemble des données générées dans la filière aquacole Calédonienne. La démarche globale de la méthodologie que nous avons établie est composée de deux principales étapes et est présentée dans la figure 5.1. La première étape consiste à modéliser le lien entre l'évolution des poids de l'espèce élevée avec des indicateurs de performance (qualité finale du produit et survie). La deuxième étape a pour objectif d'analyser la part de l'effet des conditions environnementales sur ces mêmes performances.

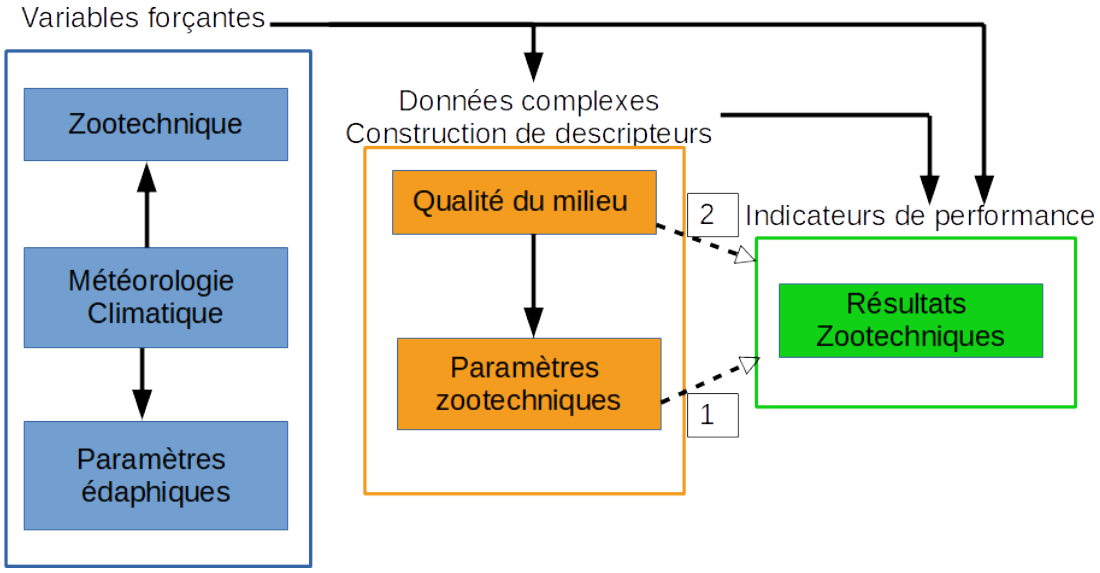

Fig. 5.1 Processus d'analyse mis en place pour l'étude de données issues d'une filière aquacole tropicale

Dans la première étape (cg. figure 5.2), nous chercherons à déterminer des groupes de croissance représentatifs des tendances à partir des données de poids relevés par les éleveurs. L'intérêt de débuter par l'analyse de ces données, est, comme énoncé dans le chapitre précédent, que leurs valeurs sont déterminantes pour le rendement des filières et que c'est l'un des paramètres les plus suivis par les éleveurs. De plus, ces données peuvent être corrélées à l'apparition de maladies, et/ou à des problèmes de qualité des produits ou encore, potentiellement, à la qualité du milieu. Pour réaliser ce groupement, nous utiliserons une méthode de clustering sur les nouveaux descripteurs construits à partir des paramètres estimés du modèle de croissance, en l'occurrence le modèle de Gompertz. Chaque groupe de tendance sera décrit et caractérisé par des données environnementales et des données de gestion des élevages. Nous verrons dans le chapitre 6 que ces groupes permettront de déterminer des périodes particulières durant l'élevage. L'étude des données de qualité du milieu, devra tenir compte de cette identification (chapitre 7). Ces périodes pertinentes sont des périodes avec un fort potentiel descriptif

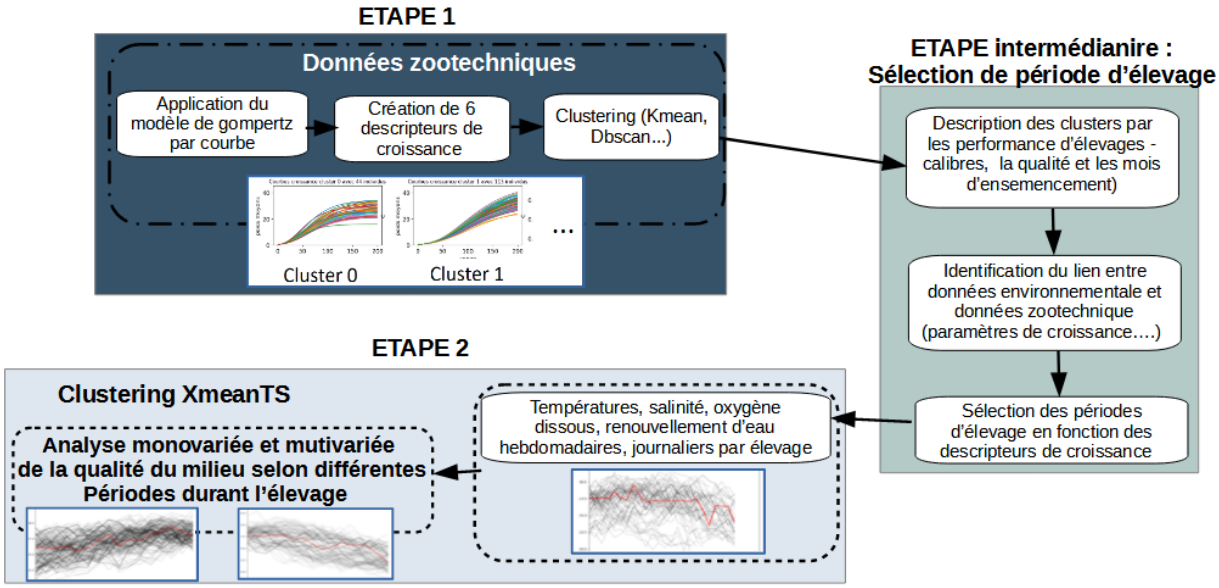

Fig. 5.2 Etapes du processus d'analyse de données de filière aquacole

et prédictif de la productivité.

Nous verrons, dans le chapitre 7, l'intérêt de croiser des données prises à différentes phases de croissance, avec l'état du milieux (i.e qualité du milieu). Nous verrons également l'avantages de nos nouvelles méthodes *Xmeans-TS* et *X-MeansMMTS* présentées dans le chapitre 4, pour analyser les données de cette filière..

5.1.1 Description général de la première étape

Comme énoncé précédemment, le prix des animaux pêchés défini par la société de transformation (SOPAC), est évalué selon divers critères, qui sont principalement zootechniques (poids moyen, taille...), lié à la qualité des animaux. Ces critères sont déterminés à l'échelle de la filière. Dans cette étape, la série temporelle des poids moyens durant l'élevage sera modélisée par une fonction de croissance adaptée qui permet d'extraire de nouveaux descripteurs zootechniques. Appliqué aux courbes de croissance des élevages étudiés, ce modèle de croissance permet de déterminer des paramètres interprétables et utilisables pour la classification non supervisée ou supervisée par d'autres données de performance et de qualité. Dans la littérature, il existe plusieurs modèles mathématiques qui permet de modéliser la croissance des animaux en aquaculture. Nous pouvons par exemple citer les modèles de *Von Bertalanfly* et de *Gompertz* [49]. Dans notre étude (chapitre 6), nous avons appliqué, aux données d'évolution du poids moyen, par élevage, le modèle de Gompertz [165], qui sera détaillé ensuite. Il a été choisi car couramment utilisé pour modéliser la croissance des crevettes. Au final 6 descripteurs de croissance seront générés à partir des courbes du modèle pour chaque élevage. Ces

nouveaux descripteurs sont par exemple l'évolution et la vitesse de convergence de la croissance.

Plusieurs méthodes de clustering seront testées sur ces 6 descripteurs de croissance, et en utilisant différentes valeurs de paramètres d'entrées propres à chacune d'entre elles (nombre de clusters *k*, densité des clusters..). Des résultats de ces tests seront comparés afin de justifier du choix des paramètres retenus pour leur affichage. Pour toutes ces méthodes de clustering, les clusters obtenus seront décrits par des données de production afin de déterminer les liens possibles entre les descripteurs de croissance et la performance des élevages. D'autres variables liées aux protocoles d'élevage pourront être testées (p.ex. le mois d'ensemencement).

Nous verrons que les résultats mettront en évidence, comme nous devions nous y attendre, un lien entre les paramètres de croissance et le mois d'ensemencement. Ce lien sera interprété par l'expert comme une relation directe entre les descripteurs de croissance avec des données environnementales, et notamment, avec l'évolution de la température de l'eau des bassins. Nous rechercherons, par conséquence, un lien entre ces différents types de données.

Rappelons que l'enjeux de cette thèse est de proposer une approche scientifique et méthodologique pour croiser des données de sources et de types variés. Ainsi des données statiques (les descripteurs de croissance) seront croisées aux données temporelles imprécises et complexes, i.e aux variables temporelles de qualité du milieu (qui peuvent être manquantes selon l' élevage).

Dans un premier temps, pour confirmer l'existence d'un lien discriminatif entre la croissance, et les défauts, des modèles de classifications supervisées multi-labels, seront testés sur ces données, en considérant les descripteurs de croissance comme attributs supervisées par les défauts (en tant que classes labélisées).

Néanmoins les paramètres zootechniques, ne peuvent pas à eux seuls expliquer les performances d'élevage. D'autres facteurs, et notamment ceux de la qualité du milieu d'élevage doivent être considérés. De ce fait, pour garantir, l'existence du lien entre les variables temporelles de qualité du milieu et la qualité de production, nous procéderons au clustering de séries temporelles de température par production. Les clusters générés seront décrits, par les défauts, et des données liées aux protocoles d'élevages (mois d'ensemencement). Nous utiliserons la méthode *(K-Shape)*.

Nous verrons (dans les chapitres 6 et 7, que l'intervalle de valeurs dans lequel évolue ces séries, est un facteur déterminant sur la productivité. Pour confirmer cela, nous affinerons les groupes de séries de température en fonction de leurs amplitudes. Les séries, par cluster seront partitionnées en fonction de l'aire médian sous leurs courbes.

Ces groupes de séries, plus 'affinés', seront décrits une nouvelle fois par les données de qualité de production, pour confirmer l'intérêt de considérer cette amplitude, dans le clustering. Cela démontera l'intérêt d'appliquer sur les données environnementales, la nouvelle méthode *X-meansTS*, qui prend en compte cet intervalle.

L'objectif enfin est de déterminer la possibilité de mettre en place un modèle prédictif des performances d'élevage en fonction de la stratégie d'ensemencement appliquée par les éleveurs, et de l'évolution de la qualité du milieu.

5.1.2 Descriptif général de la deuxième étape

Dans cette étape, les séries des variables temporelles de la qualité du milieu, seront étudiées en utilisant*X-meansTS* et *X-meansMMTS*. Les clusters obtenus, seront décris par les données de performance.

Une étude sur différents niveaux de résolutions des variables de qualité de l'eau, vise à sélectionner les résolutions en fonction de ces variables et des périodes pertinentes (avec un fort potentiel descriptif et prédictif de la productivité).

Les variables temporelles seront pour cela étudiées à différentes résolutions, par *XmeansTS* dans un premier temps. A partir de l'interprétation des résultats, les variables de qualité du milieu en fonction des périodes pertinentes, seront ensuite étudiées, par *Xmeans-MMTS* (méthode de clustering de séries temporelles multi-variés et multi-échelles).

L'analyse dans la deuxième étape, se fera donc au travers de modèles descriptifs, impliquant l'évolution de la qualité du milieu à des périodes pertinentes, les défauts et les paramètres de croissance.

L'objectif final est de déterminer un modèle de classification supervisée et non supervisée, intégrant l'ensemble des données complexes générées dans les filières aquacoles. (figure 5.3).

Les performances de *X-meansTS*, sur les données de qualité du milieu, seront régulièrement comparées à celles de l'algorithme *K-shape*. En effet, d'après la littérature, cette méthode a de très bonnes performances sur divers jeux de données en ligne [128]. Nous verrons que *X-meansTS* crée des clusters visuellement plus homogènes, sur des séries temporelles complexes (i.e avec de fortes variations). En effet, au travers de l'approche *X-meansTS*, la nouvelle mesure de dispersion, selon un seuil fixé, vise à extraire, du jeux de données, des groupes de séries avec la plus grande quantité d'informations possibles à chaque itération.

L'analyse de la qualité du milieu étant une analyse temporelle (faite avec les nouvelles méthodes crées *X-meansTS* et *X-meansMMTS*), nous chercherons à identifier des normes d'évolutions de cette qualité, et à déterminer des classes de productivité (bonne

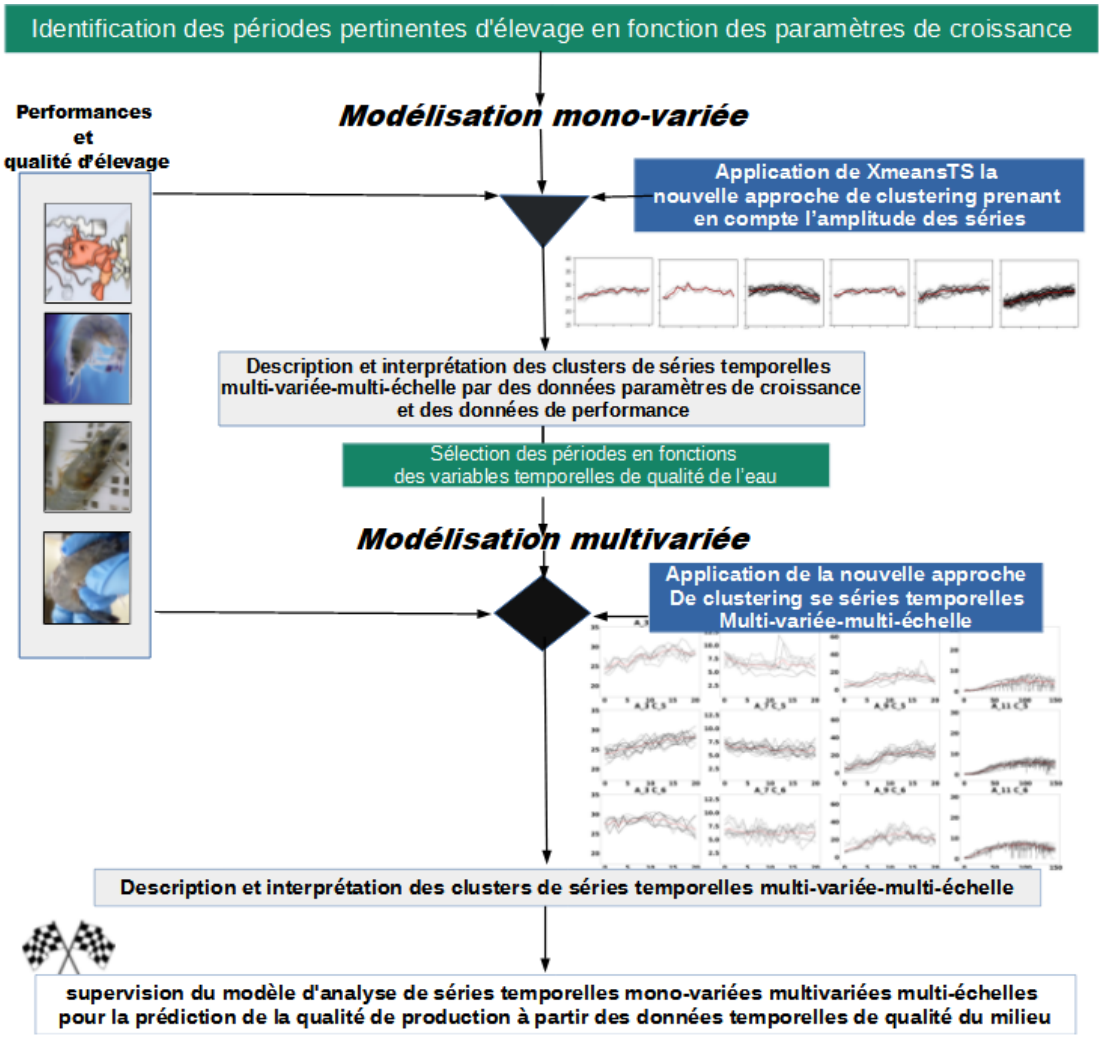

Fig. 5.3 Processus de recherche de périodes pertinentes pour une analyse multi-variée et multi-échelle

ou mauvaise survie, bon élevage...). Cela vise a créer des labels de productivités compatibles avec l'évolution de la qualité du milieu d'élevage et qui prennent en compte le climat. Pour l'analyse de la performance des clusters, nous nous intéresserons, en premier lieu, aux distributions significatives des données de productivité, par pair de clusters obtenus sur les données de qualité du milieu.

Chapitre 6

Analyse de la performance de filières aquacoles à partir des données de croissances

Ce chapitre présente en détail, la première phase de l'approche méthodologique (première étape de la méthodologie décrit dans le chapitre 5, destinée à l'analyse de la performance des filières aquacoles (survie, qualité des produits, etc.) . Comme énoncé précédemment la croissance est un indicateur primordial de qualité de production. Il permet notamment de déterminer le prix des productions dans les filières aquacoles.

Dans cette étape, les courbes de croissance sont étudiées selon un modèle de croissance mathématique adapté (mosdèle de Gompertz).

Ce modèle s'appliquera aux données temporelles du poids moyen de l'animal durant la phase d'élevage. Il permet d'extraire des paramètres de la courbe de croissance moyenne, (par élevage), afin de décrire la vitesse de croissance à différentes périodes d'élevage. Ces descripteurs serviront d'attributs pour discriminer les élevages. Les résultats du clustering ont mis en évidence des typologies de croissance. Les clusters générés seront décrits par diverses données (mois d'ensemencement, performances d'élevage..).

La figure 6.1 décrit le processus de cette analyse. Les paramètres de croissance seront dans un premier temps croisés avec des données de qualité du produit et de qualité du milieu d'élevage. Ce croisement permettra d'identifier des périodes d'ensemencement avec un potentiel descriptif (et prédictif) de la performance d'élevage en fonction de la stratégie d'ensemencement appliquée par les éleveurs. Il mettra en évidence l'intérêt de générer durant ces périodes, un clustering des variables temporelles de qualité du milieu en fonction de l'évolution de leurs formes et de leurs amplitudes dans le temps.

6.1 Effet de la croissance sur la performance

La croissance des crevettes sera modélisée par élevage selon le modèle de Gompertz qui est donné par la fonction de croissance suivante :

$$G(t) = 0.3 * \exp^{B*(1-\exp^{(-t*C)})} \tag{6.1}$$

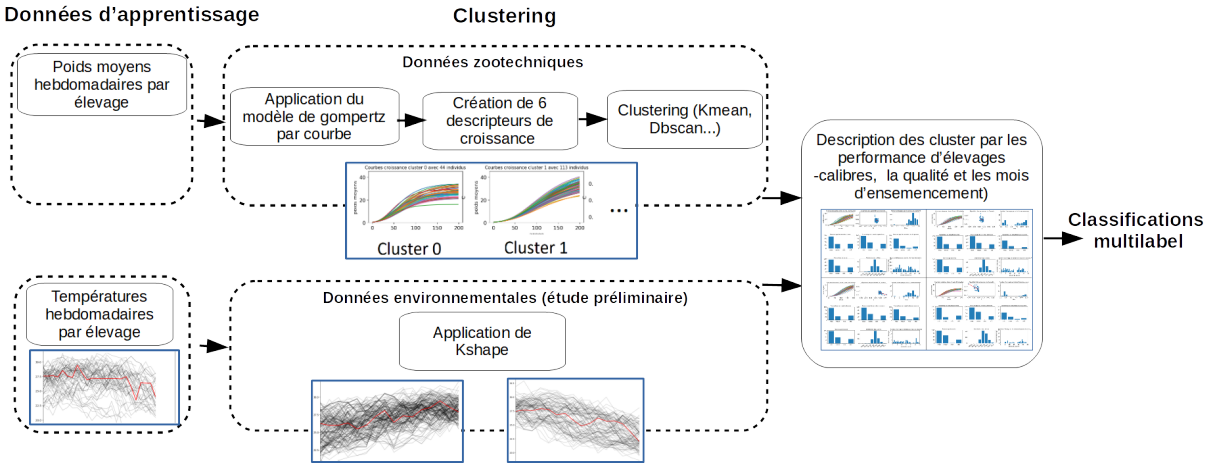

Fig. 6.1 Processus d'analyse de la première étape

La paramètre C détermine (cf. figure 6.2) l'étalement du phénomène de croissance sur l'axe des abscisses. Le paramètre B mesure la vitesse de convergence vers le poids moyen final. La courbe de croissance a une accélération plus importante en début d'élevage, et converge progressivement vers le poids final. Cette convergence se fait avec une accélération plus faible qu'en début d'élevage et crée un point d'inflexion de la courbe (cf.figure 6.3). L'exemple de la figure 6.3 montre l'arrivée de ce point au 40ème jours. Ce point s'obtient lorsque la dérivée seconde du modèle s'annule.

NB : Le point d'inflexion permettra par la suite de définir des périodes avec un fort potentiel descriptif de la performance d'élevage en fonction de la stratégie d'ensemencement appliquée par les éleveurs. La principale période correspondra à la durée d'élevage bornée par le jour d'ensemencement et l'arrivée du point d'inflexion.

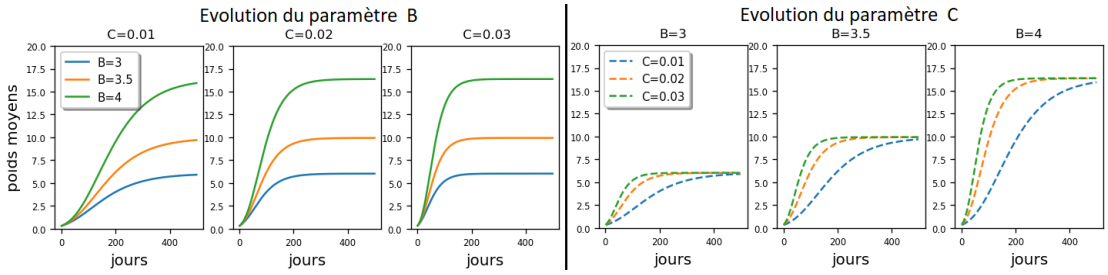

Fig. 6.2 Evolution de la croissance pour des valeurs b et C variables

Le paramètre C est appelé "taux de croissance initial" c'est à dire le taux de croissance entre le jour d'ensemencement d'un élevage (i.e. jour 0) et le jour correspondant au point d'inflexion de la courbe du modèle. L'arrivée du point d'inflexion de la croissance varie d'un élevage à l'autre. Ce point varie significativement selon les valeurs conjointes des deux paramètres B et C du modèle de Gompertz. Pour obtenir ces paramètres, la bibliothèque *easynls* (en langage R) est utilisée afin de les estimer par la

méthode des moindres carrées.

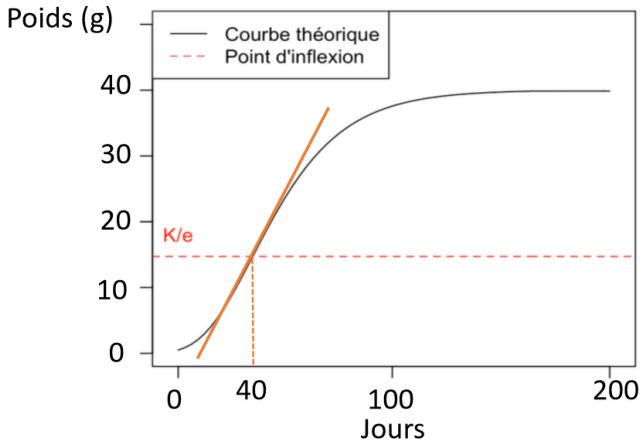

Fig. 6.3 Point d'inflexion d'une courbe de croissance

Afin d'augmenter le nombre de descripteurs pertinents en vue d'une classification non supervisée, d'autres paramètres zootechniques sont calculés. Ces paramètres sont les suivants :

- Le temps correspondant au point d'inflexion, notée P_I, déterminé par la condition nécessaire $G''(t) = 0$.
- Les temps notés G_1 et G_5 correspondant aux valeurs respectives de t vérifiant $G(t) = 1$ et $G(t) = 5$. Ces temps correspondent à la durée d'élevage pour que les animaux atteignent respectivement les poids de $1g$ et de $5g$. Ces valeurs correspondent à des étapes clés pour les éleveurs. Le poids de $1g$ correspond, généralement, à l'instant de début de suivi des variables de productions enregistrées dans la base de données *STYLIBASE*. Le poids de $5g$ correspond à la période où l'on note une forte accélération de la croissance.

La figure 6.4 compare graphiquement les modèles obtenus par la fonction de *Gompertz*, à un exemple de courbes de données réelles. On remarque une adéquation du modèle aux tendances réelles de la courbe de croissance. En effet, pour chacune des courbes de croissance étudiées, le coefficient de détermination R^2 du modèle est supérieur à $0,98$.
Au final, nous disposerons de 6 descripteurs calculés B, C, P_I, G_1, G_5 et D_e (la durée totale d'un élevage).

6.1.1 Classification non supervisée à partir de nouveaux descripteurs de croissance

Les élevages ont été regroupés selon les 6 descripteurs de croissance. Différentes méthodes de clustering ont été testées et comparées : *K-Means*, *X-Means* [129] et *db-*

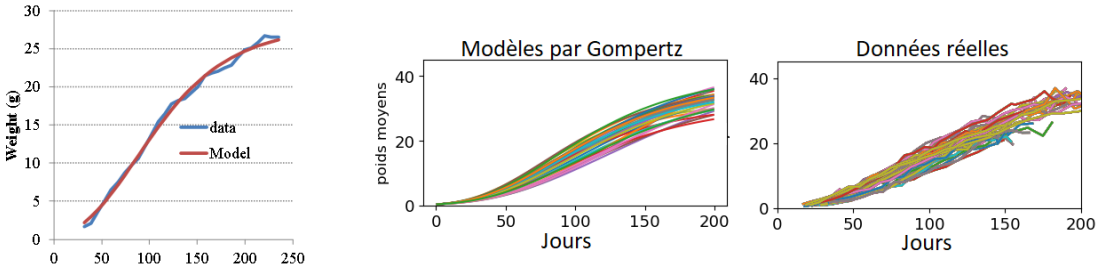

Fig. 6.4 Exemple de comparaison entre les résultats du modèle de Gompertz et les données brutes.

scan. Les 3 méthodes de clustering ont générés des clusters dont les modèles de croissance (de Gompertz) associées avaient des tendances très homogènes. Ces méthodes ont mis en évidence des typologies de croissance comparables.

Pour chaque méthode, les clusters ont été décrits par d'autres variables (mois d'élevage, défauts, groupes de poids...). Le mois d'ensemencement a été la première variable explicative de la formation des différentes typologies de croissance. Les résultats affichés concerneront principalement ceux de la méthode *K-Means*, dont le nombre k de clusters a été fixé avec les experts, après interprétation des résultats de plusieurs tests avec différentes valeurs de nombre de clusters paramétrés. L'interprétation des experts sur les résultats générés en faisant varier k de 2 à 10 a permis de fixer le nombre de clusters pertinent pour mettre en évidence des informations intéressantes concernant les typologies d'élevage. Il a été retenu qu'un regroupement des courbes de croissance en 5 clusters permet d'avoir, un nombre représentatif d'individus par clusters. De plus des tests comparatifs seront présentés entre les résultats de ces 5 clusters par *K-means*, avec les clusters générés par la méthode de clustering *X-means* [129]. La description des clusters de ces deux méthodes, par différentes variables (défauts, paramètres de croissance...) montrera que la variable explicative prépondérante reste, dans les deux cas, le mois d'ensemencement. Cette information servira dans l'analyse de la qualité d'eau, qui sera détaillée ensuite dans le chapitre 7, car le mois d'ensemencement détermine la température d'eau des bassins.

La figure 6.5 montre le résultat d'un clustering des 6 descripteurs de croissance par la méthode *K-means*. Des typologies de croissances associées à chaque cluster sont observables d'après la forme des courbes (de poids) de leurs individus (i.e élevages). Ces typologies caractérisent 5 pratiques d'élevages de la filière, qui peuvent être décrites par des variables forçantes (mois d'ensemencement, âge du bassin...).

La méthode *DBscan* a fournit des résultats presque similaires à la méthode *K-means* en regroupant les clusters selon les typologies de croissances représentés dans la figure 6.5.

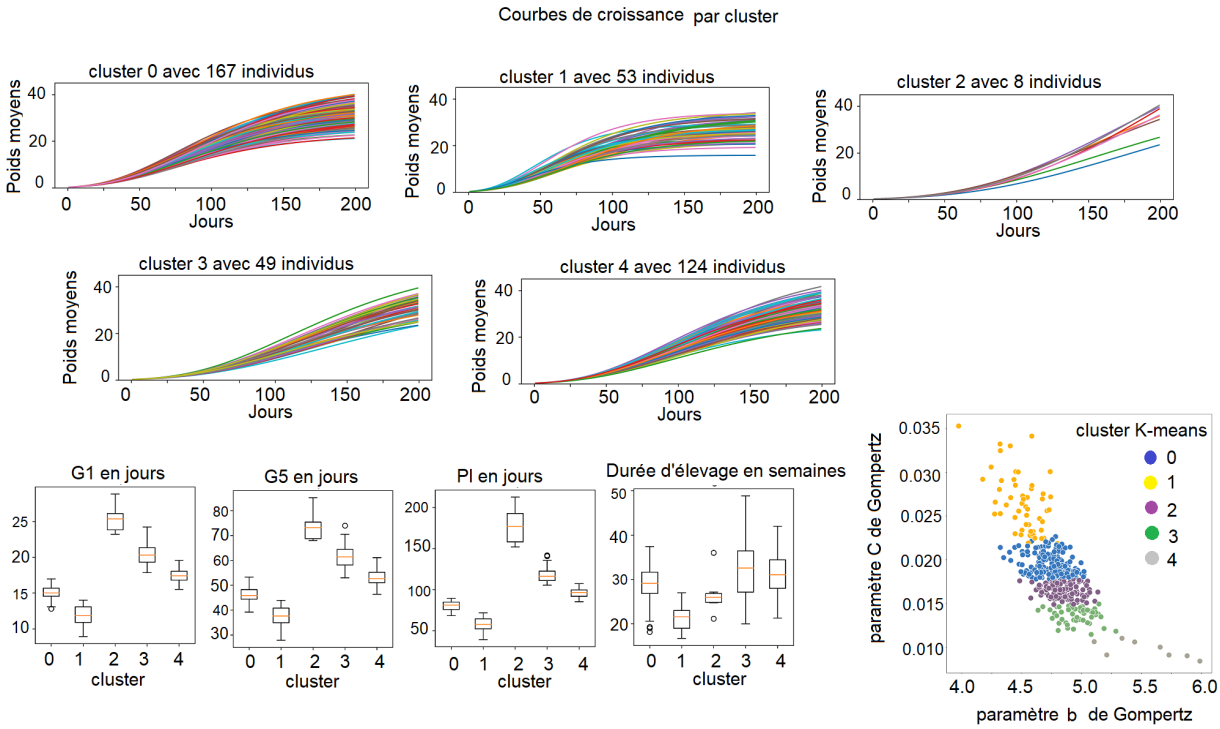

Fig. 6.5 Principaux groupes obtenus par les descripteurs de croissance

6.1.1.1 Description des 5 clusters de croissance générés par *K-means*

La figure 6.6 décrit les clusters selon l'âge des bassins et les périodes d'ensemencement des élevages. La description des clusters selon l'âge moyen des bassins vise à déterminer si l'exploitation répétée du terrain impacterait ou non la qualité de l'espèce, et notamment sa croissance, car elle vit principalement au fond des bassins. Néanmoins, avec les 6 descripteurs de croissances, les distributions des valeurs d'âges des bassins et difficilement interprétable contrairement aux distributions des mois d'ensemencement entre ces clusters. Par exemple, le cluster 1 correspond à un groupe d'élevages ensemencés en début d'année (cf. figure 6.5), leurs courbes de croissances convergent rapidement vers leur valeur finale. La Nouvelle-Calédonie étant située dans l'hémisphère Sud, ces élevages débutent pendant la période chaude. Il y a deux principales périodes : la période 'chaude' et la période fraîche qui débute au mois de juin. En raison de ce climat, la différence de distribution des valeurs des descripteurs entre le cluster 1 et les clusters 3 et 4 peut être expliquée par une température saisonnière plus élevée pour le cluster 1 que pour les autres clusters. Ces résultats sont confirmés par le cluster 2. En effet, bien que ce cluster soit marqué par un faible nombre d'individus, ses élevages sont ensemencés entre les mois de juin et août correspondant à la période fraîche. Il montre une survie plus faible et un taux de têtes éclatées plus élevé (par rapport aux autres clusters) et donc les performances d'élevage faibles (figure 6.7).

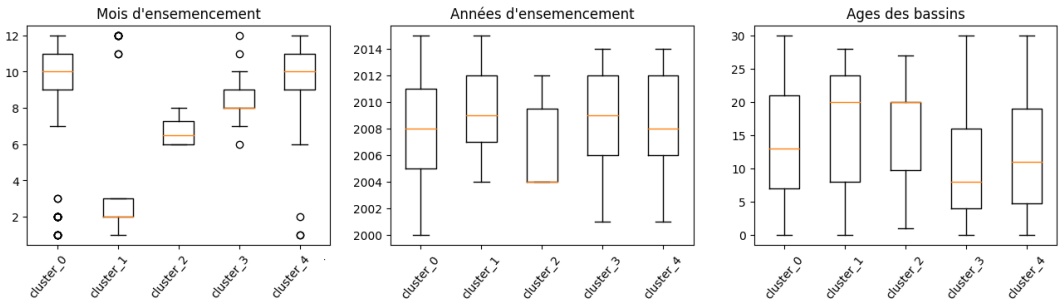

Fig. 6.6 Description des clusters par des variables temporelles

La figure 6.7 montre la répartition moyenne des défauts et de la survie dans chacun des clusters. La figure 6.8 présente la répartition des différents calibres des élevages par cluster. Selon ces deux figures les distributions des défauts, et des calibres, par cluster permettent difficilement d'interpréter les typologies de croissance.

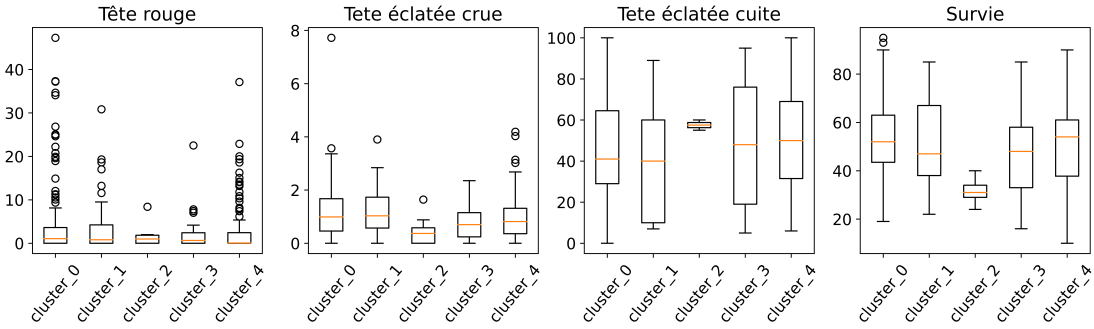

Fig. 6.7 Qualité des groupes d'élevages

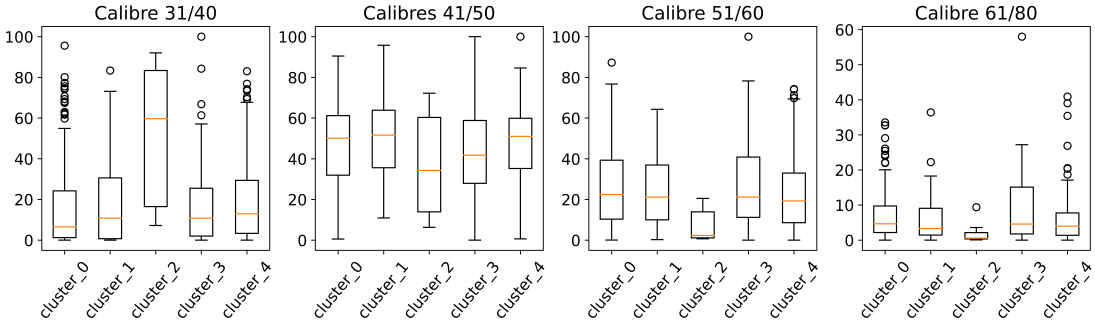

Fig. 6.8 Calibres des groupes d'élevages

Nous nous sommes intéressés aux données de qualités d'élevage (i.e les calibres, les défauts, la survie....). L' analyse de la variance *ANOVA* a été utilisée pour comparer les distribution de différentes variables de qualité d'élevage par paire de clusters. La figure 6.9 montre, comme exemple, des matrices de p-valeurs obtenus par *ANOVA* en

considérant les données par paire de cluster, associés aux défauts *tête chargées* et *patte verte*, au calibre *31/40* et enfin à la variable du mois d'ensemencement. Selon ces résultats la matrice de p-valeurs du mois d'ensemencement contient les p-valeurs les plus faibles, et comparativement aux autres variables, elle a d'avantage de p-valeurs inférieur à 0.05. Elle est la variable explicative prépondérante.

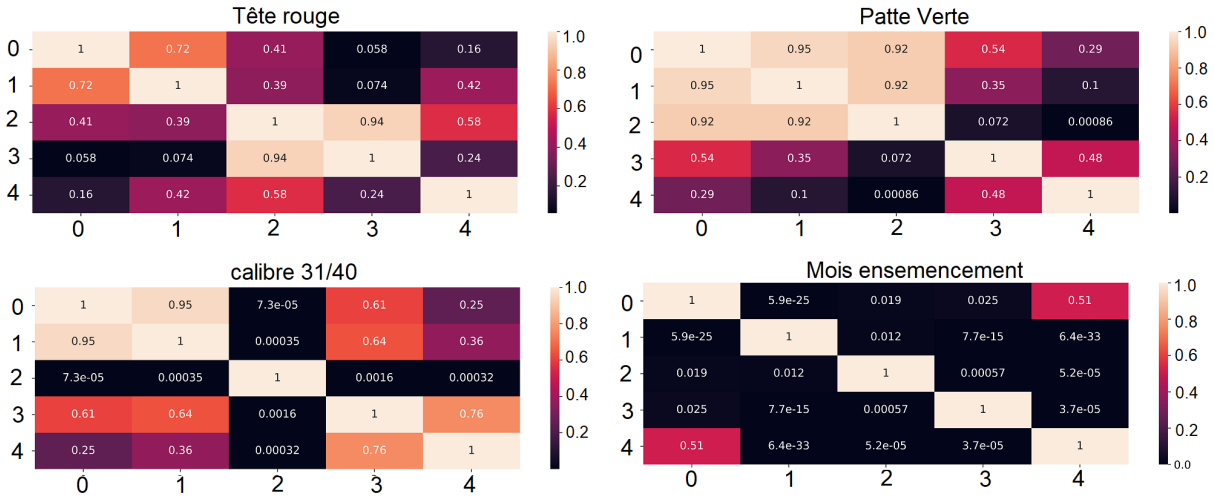

Fig. 6.9 P-Valeurs obtenus sur des données de qualité d'élevage, par pair de clusters

6.1.1.2 Validation du nombre de clusters ($k = 5$)

Pour valider le nombre pertinent de clusters à prendre en compte, nous avons utilisé et appliqué la méthode *X-means* et *DBScan* qui permettent de fournir automatiquement le nombre de clusters. Nous avons choisi ici de ne discuter que les résultats de *X-Means*, *DBSCAN* fourni les mêmes résultats. La méthode *X-Means* a été appliquée et les résultats ont été comparés aux résultats fournis par *K-means* avec $k = 5$. *X-Means* a généré automatiquement 11 clusters qui ont été analysés et comparés aux k clusters générés par $k - Means$. La figure 6.10 montre les 11 clusters affichés dans le plan des descripteurs B et C avec une analyse descriptive, via des boxplots, pour se rendre compte des variabilités des autres descripteurs ($G1$, $G5$, PI et De) par cluster.

Pour vérifier quelles sont les variables explicatives (données zootechniques comme les paramètres de Gompertz, les défauts..) qui décrivent de manière discriminatoire les 11 clusters obtenus, nous avons procédé à une analyse de la variance *ANOVA* pour comparer les clusters 2 à 2 par variable explicative. Cette analyse a confirmé (encore) que la variable d'ensemencement est la seule qui décrit statistiquement le mieux les clusters.

La matrice 6.11 montre les p-valeurs par paire de clusters, obtenus par *X-means*, pour la variable "mois d'ensemencement". Pour confirmer le choix d'un k=5 pour

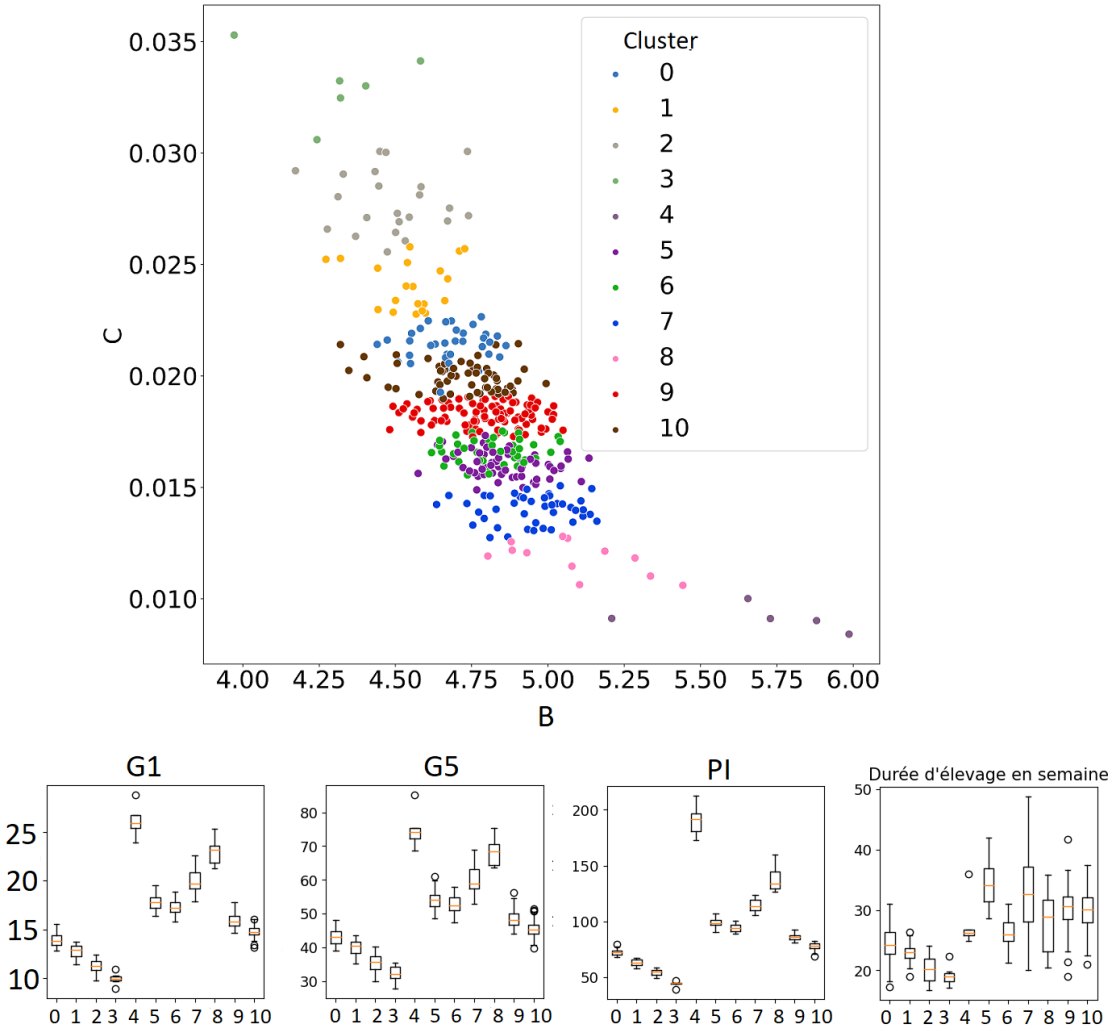

Fig. 6.10 Description des 11 clusters obtenus par *X-means* sur les descripteurs de croissances

obtenir les 5 typologies de croissances par *K-means*, nous verrons qu'il est possible de regrouper les 11 clusters de *X-means* à 5 clusters selon les p-valeurs de la matrice 6.11.

En effet, la figure 6.12, présente la distribution des valeurs de mois d'ensemencement par cluster (fourni par *X-Means*). Quatre groupes de clusters ont les mêmes tendances (cf. figure 6.12 et figure 6.13) : $Gp1 = \{ 1, 2, 3 \}$, $Gp2 = \{ 5, 6 \}$, $Gp3 = \{ 7, 8 \}$ et $Gp4 = \{ 9, 10 \}$. les clusters de chaque groupe sont ensemencés pratiquement aux mêmes périodes de l'année. La matrice de p-valeurs, présentée dans la figure 6.11 confirme le rejet de l'hypothèse selon laquelle, par groupe, il y une différence significative de la moyenne des mois d'ensemencement des élevages. En effet à l'intérieur de ces groupes, les p-valeurs par paire de clusters sont supérieures à 0.05. De plus, les courbes des séries des groupes précités, ont des tendances visuellement proches (6.13).

Les périodes d'ensemencement du cluster 0 obtenu par la méthode *X-means*, peuvent être réparties dans les 4 groupes de clusters. Celles du groupe de clusters 4 sont

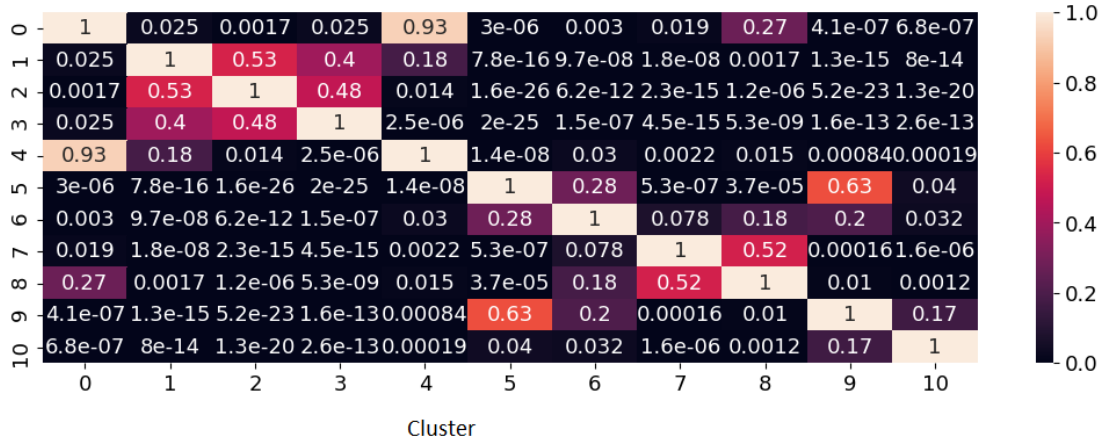

Fig. 6.11 Matrice de p-valeurs obtenues sur la variable explicative 'mois d'ensemencement'

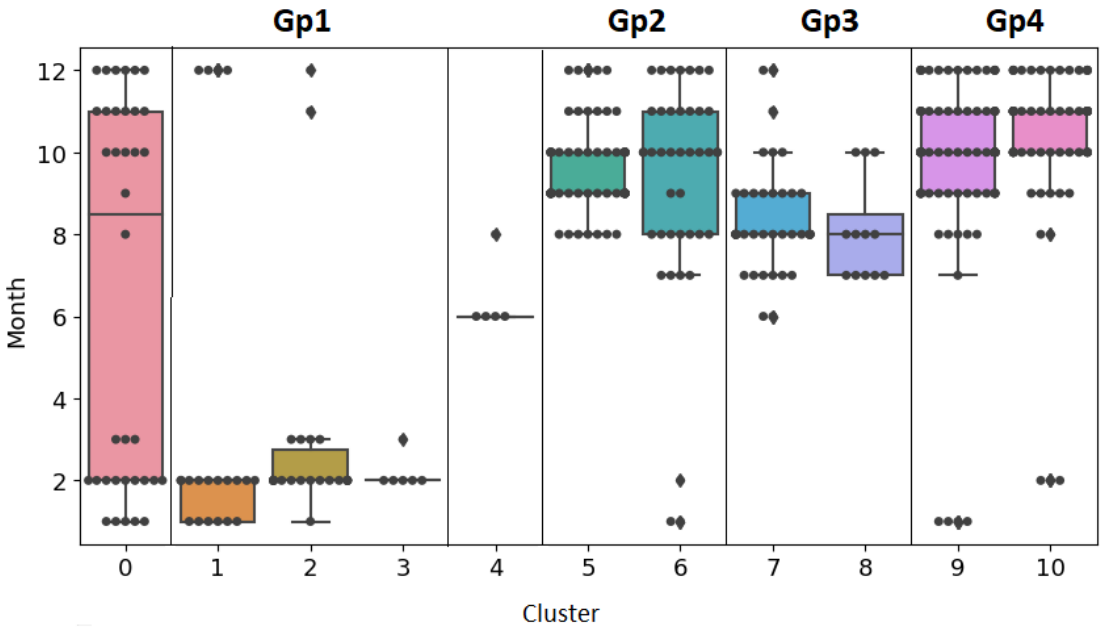

Fig. 6.12 Mois d'ensemencement par cluster obtenu avec la méthode x-means

comparables aux séries du groupe de clusters 2 obtenue par la méthode *X-Means*. Ce sont des élevages ensemencés en périodes fraîches. Le choix du nombre k de clusters représentatif de la filière, fixé à 5, dans la méthode *K-Means* est donc justifiable (statistiquement), en fonction de la variable explicative prédominante qu'est le mois d'ensemencement.

Les différents clusters obtenus (par *K-means* et *X-means*) ont relevé un lien entre croissance et mois d'ensemencement ce qui s'explique par une influence de la température d'eau des bassins sur la croissance des espèces élevées comme l'ont montré [74]. L'hypothèse qu'il y a une corrélation entre les défauts, la croissance, le mois d'ensemencement reste à vérifier. Pour ce faire, une analyse supervisée par ces vari-

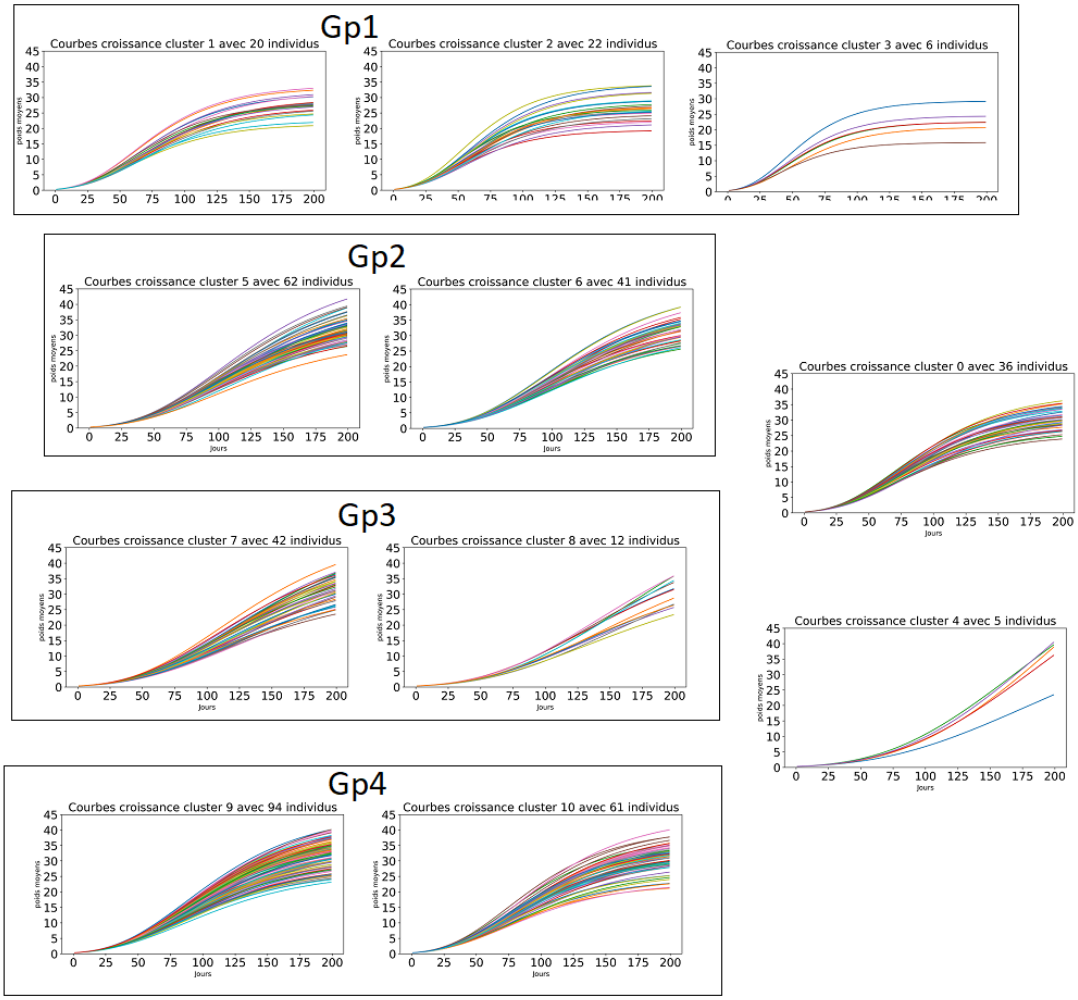

Fig. 6.13 Principaux clusters obtenus par les descripteurs de croissance avec la méthode x-means

ables explicatives a été réalisée par l'utilisation des méthodes de classifications multi-labels en considérant comme attributs les descripteurs de croissance et comme labels les différents défauts, la survie et les calibres.

6.1.2 Classification multi-label, des paramètres de croissance supervisée par les données de performance d'élevage.

Dans cette section, nous avons étudié le lien entre les 6 descripteurs construits en section 6.1 et les données de qualité. Pour cela, nous avons opté pour une méthode de classification supervisée en considérant ces données de qualité comme variables cibles. En effet, La qualité du produit à chaque pêche est estimée, à partir du pourcentage des différents défauts relevés par l'usine de conditionnement. Calibres et défauts deviendront donc les principales variables à prédire, i.e. des labels ou classes pour un problème de classification supervisée en plus de la survie qui reste un facteur très im-

portant économiquement. La particularité dans notre cas est qu'on est dans une situation de classification supervisée multi-labels et multi-classes (puisqu'on est en présence de plusieurs variables à prédire).

6.1.2.1 Approches existantes de classification supervisée multi-labels

Comme énoncé dans le chapitre 2, il existe dans la littérature de nombreuses approches concernant la classification multi-label [6]. Dans [169], les auteurs les regroupent en deux grandes catégories. La première catégorie englobe les méthodes qui adaptent les algorithmes monolabel pour traiter directement des données multi-label. La seconde catégorie fait référence aux méthodes qui transforment un problème mono-label en un problème multi-label. Dans [184] les auteurs différencient également les méthodes selon qu'elles considèrent ou non les dépendances possibles entre les labels.

Dans l'apprentissage multi-labels, les performances prédictives optimales sont obtenues par des méthodes considérant explicitement les dépendances possibles entre les labels [33]. Les notions de corrélation et de dépendance entre les labels ont été discutées [34]. Par exemple, les méthodes de type *chain* (CC) comme la méthode *Probabilistique Classifier Chain* (PCC) ou encore *Ensemble Classifier Chain* (ECC) [32], proposée par le même auteur, sont des techniques d'apprentissage en chaîne. Elle détermine la relation que possède chaque label avec les attributs en lui associant, lors de son intégration dans le modèle, un coefficient lié à sa loi marginale calculée par rapport aux lois de probabilités des labels intégrés.

L'avantage des méthodes de type classifieur chain *CC* réside dans la performance du temps d'exécution de la phase d'apprentissage et également dans la formulation des corrélations entre les labels. Cependant, dans ces méthodes, les labels (par individu) sont analysés dans un ordre défini aléatoirement, le résultat en dépend fortement ce qui est une faiblesse [6] à prendre en compte.

6.1.2.2 Classification supervisée multi-labels des données aquacoles

Nous avons testé les deux approches de classification supervisée multi-labels sur les données en considérant les 6 descripteurs extraits du modèle de Gompertz. La première approche est représentée par les *classifiers chains* qui prennent en compte la dépendance entre les labels sélectionnés (nous avons choisi les techniques *PCC* et *ECC*). La seconde approche concerne les techniques ne prenant pas en compte cette dépendance (comme par exemple la méthode*Binary Relevance* (BR)). Chaque technique étant une adaptation des méthodes de classification mono-label, nous avons donc utilisés en entrée, les classifieurs mono-label suivants : *Decision trees*, *Random Forest*, *Nearest neightbourg* et *SVM*.

Concernant les labels, nous considérons des données de qualités (calibre, défaut)

et de quantité (liée à la survie). Parmi les défauts relevés par la SOPAC, nous nous concentrons sur 3 défauts visibles au niveau de la tête de la crevette qui contient les principaux organes internes (coeur, cerveau, estomac...). Ces 3 défauts sur la tête sont ciblés dans notre étude car ils ont une influence importante sur le déclassement des produits de la filière. Ces défauts sont :

- $d1$: "tête rouge"
- $d2$:"tête éclatée crue"
- $d3$:"tête éclatée cuite"

Les calibres choisis en tant que labels sont les calibres les plus produits dans la filière : $31/40$, $41/50$, $51/60$ et $61/80$. Nous considérerons également la survie comme variable cible pour estimer les performances d'élevage.

Les variables de qualité à prédire (variables de type quantitatives) sélectionnées précédemment, ont été discrétisées en classes (intervalles) de fréquences égales pour les coder en tant que classes d'apprentissage pour les classifieurs multi-labels.

Le tableau 6.1 présente les performances de plusieurs classifieurs multi-labels. Globalement les méthodes considérant les dépendances entre les labels (*PCC*, *ECC*) ont une meilleure performance que la méthode *BR* (qui effectue un apprentissage de manière indépendante sur chaque cible). La méthode ECC *Ensemble classifier chain* qui construit plusieurs classificateurs en chaîne par un ordre d'étiquettes aléatoires, a la performance la plus élevée. Cette première analyse, met en avant l'hypothèse d'une potentielle dépendance entre les paramètres de Gompertz, et l'ensemble des cibles utilisées i.e les défauts et les calibres. (figure 6.16).

L'interprétation de ces résultats n'est pas intuitive, car ces méthodes multi-labels ne restituent pas graphiquement de modèles (d'interprétation) de l'apprentissage. En effet, par exemple les modèles mono-labels, basés sur une structure telle que les arbres de décision fournissent une représentation graphique qui hiérarchise l'influence des attributs sur la cible. Dans l'apprentissage multi-label, l'approche *Ensemble classifier chain*, qui peut utiliser l'arbre de décision comme méthode de base, met à jour le modèle d'apprentissage en incluant les cibles de manière itérative dans l'apprentissage après avoir été supervisé par chacune d'elle (de manière itérative). Dans ce dernier cas il y a donc bien un apprentissage basé sur la recherche d'une dépendance entre les attributs et les cibles.

Par conséquent, les résultats obtenus (avec les approches multi-labelles) montrent qu'il y a une corrélation (qui peut être complexe) globale entre les variables suivantes : le mois d'ensemencement, la croissance initiale, la vitesse de convergence vers le poids final, les défauts et les calibres. Or comme énoncé précédemment, la température (de

Classifieur Multi-label	Classifieur de base	Rappel	Précision
ProbabilisticClassifierChain	RDMForest	0,99	0,97
EnsembleClassifierChain	RDMForest	0,99	0,95
BinaryRelevance	RDMForest	0,97	0,96
ProbabilisticClassifierChain	KNN	0,78	0,78
EnsembleClassifierChain	KNN	0,78	0,78
BinaryRelevance	KNN	0,78	0,77
ProbabilisticClassifierChain	DecTree	0,74	0,76
EnsembleClassifierChain	DecTree	0,73	0,76
BinaryRelevance	DecTree	0,73	0,76

Table 6.1 Performances des classifieurs multi-labels sur les données de qualité de production.

l'eau) est le facteur principalement lié aux typologies de croissance de la filière. De ce fait, des groupes obtenus en fonction de clusters de croissance seront, par la suite, décris par des données temporelles de qualité du milieu et notamment par les séries de température d'eau.

6.2 Classification non supervisée des variables temporelles d'environnement et de gestion à partir des performances de l'ensemble des élevages

L'analyse que nous allons effectué dans cette section concerne l'identification des tendances des variables temporelles de qualité du milieu (température, renouvellement d'eau, salinité..) en fonction des variables de croissance. L'idée est de représenter chaque cluster (en nombre de 5), généré grâce aux descripteurs extraits des modèles de croissance, par ses séries temporelles des variables de qualité du milieu (température, oxygène, salinité de l'eau, etc.).

Dans la figure 6.14, les séries de température de l'eau, sont affichées en fonction des 5 groupes de croissance des élevages identifiés dans la section précédente.

Des tendances sont observables, dans ces séries de températures en fonction des clusters de croissance 1,2 et 3. Néanmoins le faible nombre d'individus des clusters 1 et 2 ne nous permettent pas de présenter ces résultats, comme des résultats représentatifs de la qualité du milieu des élevages de la filière. Ces élevages doivent être considérés comme exceptionnels. Aucune tendance n'est perceptible pour les autres variables environnementales telles que l'oxygène dissous et la salinité (figure 6.15). Il y a aussi des tendances observables sur les variables de gestion tels que le renouvellement d'eau et l'alimentation Néanmoins la température est la variable présentant le plus de tendances hétérogènes entre cluster.

NB : Les quartiles ont été calculés pour chacune des variables de performance

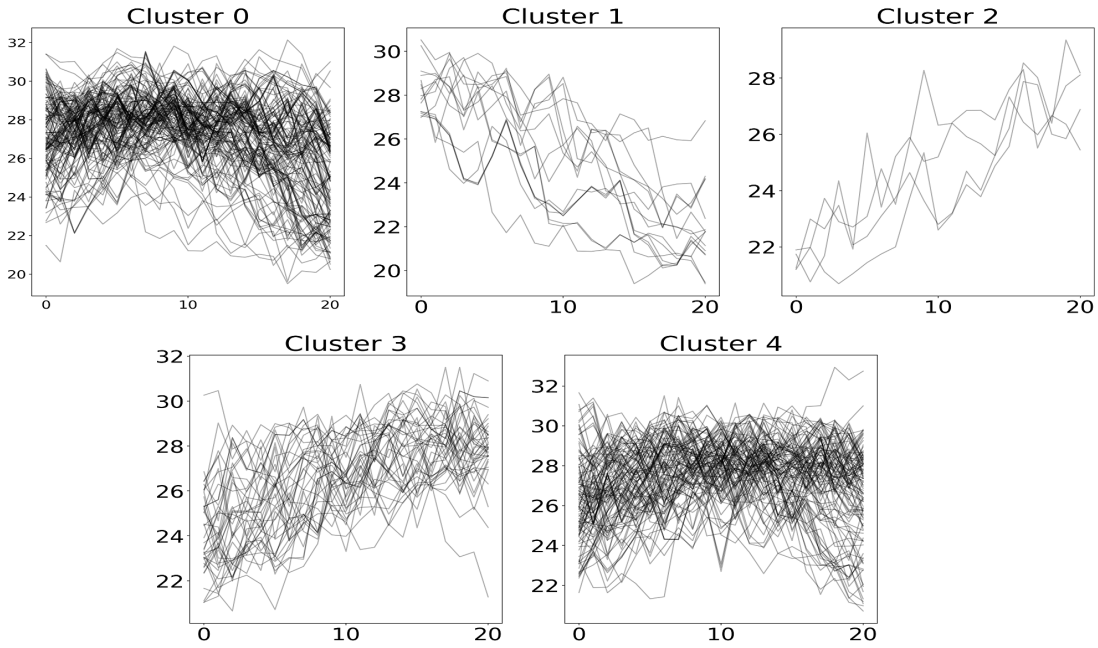

Fig. 6.14 Séries temporelles de température en fonction des clusters de croissance

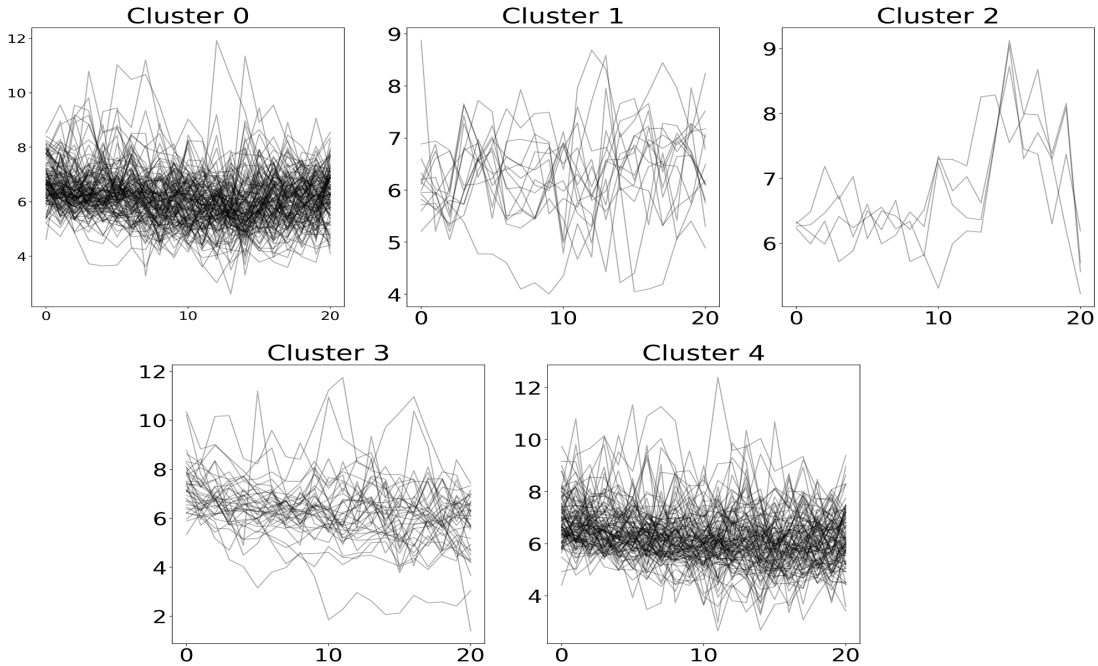

Fig. 6.15 Séries temporelles d'oxygène dissous en fonction des clusters de croissance

d'élevage (défauts, survie, données zootechniques...). Pour chacune de ces variables, leurs groupes (classes) de quartiles ont été décris par les variables temporelles de qualité d'eau. Visuellement, les séries temporelles par classe n'étaient pas homogènes. Il y a une difficulté à obtenir des groupes de séries temporelles homogènes à partir des classes de fréquences obtenues sur les données de performances d'élevage.

D'après la classification supervisée multi-label précédente, un lien existerait entre les paramètres de croissance, la survie et les défauts. Avec la description des clusters de croissance, par les séries de température, on peut envisager un potentiel lien entre les paramètres de Gompertz et la température de l'eau des bassins.

Il y aurait donc un lien entre paramètres zootechniques et données de performance d'élevage telle que la survie. Il y aurait également un impacte de la qualité du milieu (température, renouvellement d'eau ...) sur les paramètres zootechniques (paramètres de croissance). Dans l'autre sens, ces paramètres de croissance peuvent aussi impacter la qualité du milieu, en raison des matières organiques produites par les crevettes au cours du grossissement. Les paramètres zootechniques sont des indicateurs primordiaux pour la rentabilité de la filière. Mais ils ne peuvent pas être considérés, dans un modèle d'apprentissage, comme les seuls indicateurs liés à la performance. Avant de poursuivre sur une recherche de modèles prenant en compte toutes les données aquacoles, le lien potentiel est à définir entre la qualité du milieu et la performance des élevages, telle que la survie (figure 6.16).

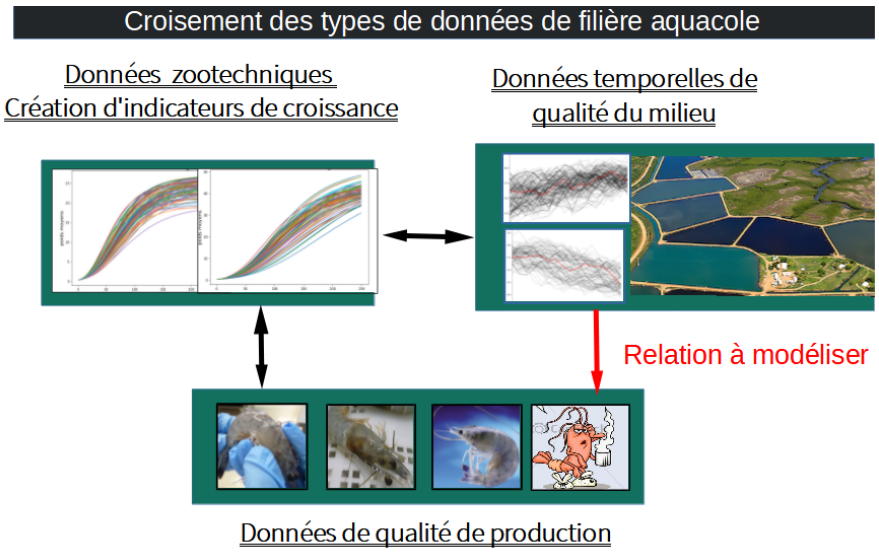

Fig. 6.16 Croisement de données générées dans l'aquaculture

Nous nous servirons, par la suite de modèles descriptifs, où les attributs seront les données de performances des élevages. Les clusters générés seront décrits par les séries temporelles de qualité du milieu.

6.2.1 Descriptions des clusters de performance par les variables temporelles de qualité du milieu

Les données de performances d'élevages ont été regroupées, à partir des méthodes *K-means* et *Dbscan*. Ces données concernent les défauts sur la têtes (”tête rouge”, ”tête éclatée cru, tête éclatée cuite) et la survie. Les clusters obtenus seront ensuite décrits par la température. En effet comparativement aux autres variables temporelles de qualité du milieu, les séries de températures ont des tendances plus homogènes, à l'intérieur des clusters générés par les méthodes *K-means* et *Dbscan*. Les résultats des clustering étant assez similaires pour ces deux approches, ceux de la méthode *K-means* seront discutés ensuite.

Plusieurs tests ont été effectués par la méthode *kmeans* en faisant varier le nombre de clusters k. Nous afficherons comme exemple, avec un nombre de clusters $k = 8$, afin de relever visuellement l'existence du lien entre les clusters de performance d'élevage et la qualité du milieu.

D'après la figure 6.18 il y a des tendances dans les séries de température, liés aux clusters de performance. Ces tendances sont comparables à celles des clusters de croissance. Par exemple, ces tendances sont observables pour les séries, des groupes 4 (tendance croissante) et 5 (tendance décroissante) de la figure 6.18, liées aux clusters de données de performance (obtenu par *K-means*). On remarque que ces groupes 4 et 5 ont respectivement les valeurs de survies les plus faibles et les plus élevées (6.17). De plus, entre ces clusters, et en début d'élevage, les températures sont comprises dans des intervalles assez distantes. L'hypothèse posée ici, et que nous vérifierons ensuite, est qu'en fonction de l'intervalle des valeurs de température, dans les premiers jours d'élevage, la croissance et la survie des espèces pourraient corrélées de manière positive. Cet impact sera modélisée dans le chapitre 7 suivant.

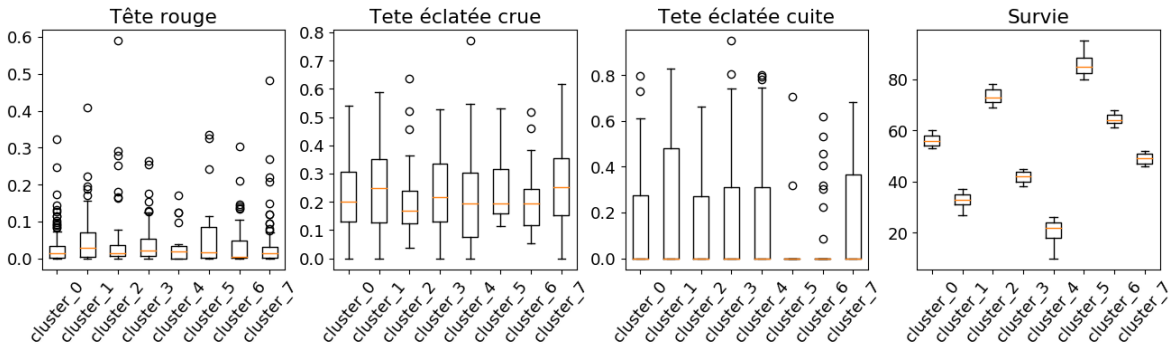

Fig. 6.17 profile de température en fonction des clusters de performance

Afin de confirmer cette hypothèse, une première analyse concerne l'évolution temporelle de température de l'eau des bassins. Ces séries ont été analysées, sur la to-

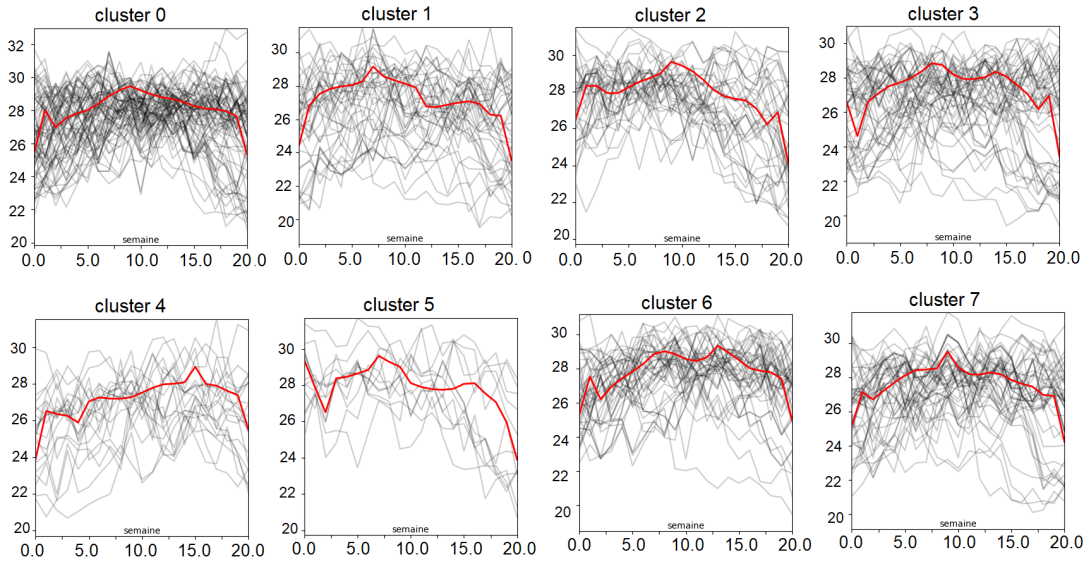

Fig. 6.18 Profil de température en fonction des clusters de performance

talité de la période des élevages, par une méthode de clustering adaptée aux séries temporelles.

La méthode *k-Shape*, développée pour clusteriser les séries temporelles avec une distance appropriée, par exemple, permet de paramétrer le nombre souhaité de cluster (k). Dans la suite nous appliquerons *K-Shape* aux séries temporelles par variable de qualité du milieu (température, renouvellement d'eau...). Nous étudierons l'influence de l'intervalle d'amplitude des séries, sur la performance d'élevage et notamment la survie. Pour cela, un nombre relativement faible (2 ou 3) de clusters sera généré par la méthode *K-Shape*. Dans les clusters, les séries seront (re)-partitionnées par rapport à une valeur médiane, afin d'obtenir des groupes avec des séries évoluant sur un intervalle d'amplitudes plus réduit. Les distributions des données de performance des élevages (relatifs aux séries) seront étudiées en fonction de ce nouveau partitionnement.

6.2.2 Analyse des tendances des séries temporelles de température par *K-shape*

Une première analyse a été réalisée pour étudier l'impact des tendances des séries temporelles de température de l'eau des bassins, sur la survie. On considère chaque série temporelle prise sur 20 premières semaines de la période de grossissement. Pour rappel, cette période varie de 4 à 6 mois en fonction des élevages. Nous appliquerons la méthode de clustering *K-shape*.

Pour rappel, la méthode *k-Shape* impose de fixer le nombre (k) de clusters comme résultat final. Il s'agit à partir d'un nombre de clusters fixé, de déterminer si dans chaque cluster de séries de température, les séries avec des températures élevées obtiennent des survies plus élevées ou non. Pour cela les séries par clusters sont partitionnées

selon une valeur médiane des aires sous leur courbe. Cette séparation d'individus par cluster de température, vise à valider le fait que, quels que soient les clusters de séries de température, les individus dont les valeurs sont plus élevées, ont une survie plus importante. L'utilisation d'une méthode de clustering existante pour séparer en amont les séries, selon l'évolution de leurs formes, permet d'obtenir dans un premier temps des groupes de séries avec des tendances homogènes. Ensuite, l'affinement des séries, décrites par des données de performance, permet de relever si des clusters de séries ayant des évolutions de formes analogues, ont des performances comparables, si on les sépare selon leurs amplitudes. Ainsi, avec $k = 2$, *K-shape* regroupe les courbes de température en deux tendances distinctes (croissante et décroissante) qui sont représentatives des variations de température que les crevettes subissent pendant l'élevage en fonction du mois d'ensemencement. Ces tendances ont relevées une survie plus importante en moyenne lorsque la température est plus élevée (figure 6.19). Afin de confirmer ce résultat dans chacun de ces clusters, générés par *K-Shape*, les individus seront encore séparés. La séparation sera faite en fonction des aires sous la courbe des séries de température. Ces aires sont obtenues à partir des valeurs d'intégrales des séries données par $\int_a^b f(x)\,\mathrm{d}x$. A partir des aires sous les courbes de l'ensemble des séries, l'aire médian est calculé. Et les séries, par cluster, sont séparées par rapport à une aire médiane, obtenue en considérant les aires sous les courbes des séries (d'un cluster). De la même manière, la séparation a été réalisée en tenant compte des autres quartiles des aires.

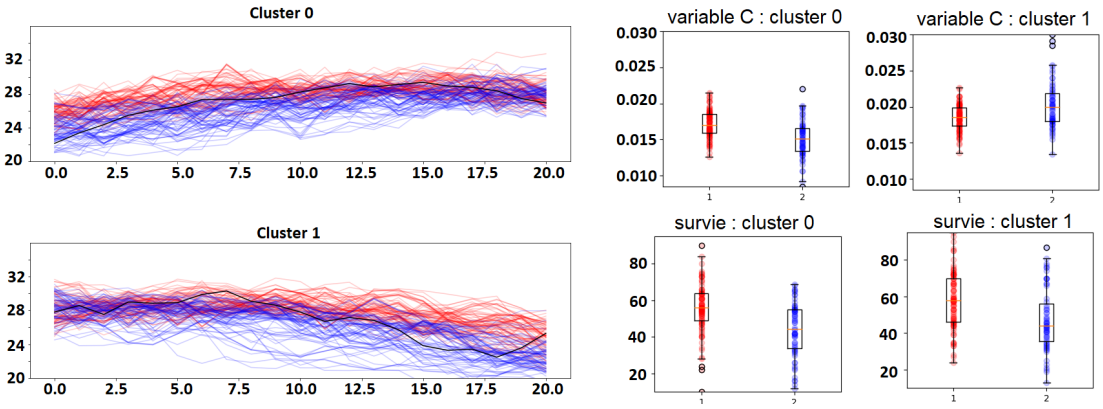

Fig. 6.19 Clustering des séries temporelles de température par la méthode *k-shape*

Nous avons donc, pour chaque cluster montré en figure 6.19, affiché en rouge les séries de températures les plus élevées i.e. au dessus de l'aire médian (2ème quartile d'aire), dans le cluster. Et en bleu les températures les plus fraîches i.e. en dessous de l'aire médian. Pour les courbes rouges, des deux clusters, la survie est en moyenne supérieure à 55% et est meilleure que la survie des courbes bleues. Cependant, nous

pouvons remarquer que le taux de survie est plus élevé dans le cluster dont les températures sont d'une part décroissantes, et d'autre part plus élevées en début d'élevage. Ce cluster présente une distribution des taux de survie avec un 3ème quartile à 70% contre 60% pour le second cluster (de tendance croissance).

Pour confirmer l'utilité d'une analyse des séries selon un affinement sur l'amplitude, un nombre de clusters plus important est crée par la méthode *K-Shape*. Cela entraîne un affinement des séries des clusters autour du représentant. Ces clusters initialement générés par *K-Shape* sont ensuite ré-affinés, comme précédemment, selon les quartiles calculés à partir des aires sous les courbes des séries. De brèves remarques seront faites sur la description des clusters. Des hypothèses seront émises et le chapitre confirmera certains d'eux, par une analyse mono-variée et multi-variée, sur des périodes particulières (en fonction du point d'inflexion). La figure 6.20 affiche les 10 clusters obtenus par *KShape*, avec en rouge (resp. bleu), pour chaque cluster, les séries dont l'aire sous la courbe est supérieure (resp. inférieur) à l'aire médiane.

Dans la plupart des cas, la survie est plus importante lorsque la température est plus élevée (figure 6.21). Visuellement, cette remarque se vérifie plus souvent pour les premières semaines d'élevages.

La distribution de la vitesse de croissante, par cluster d'évolution temporelle de température d'eau, est cohérente avec les résultats du clustering sur les paramètres de croissance de Gompertz. Dans la plupart des cas, plus la température est élevée et plus le paramètre de Gompertz C (le taux de croissance initial) est élevé. Il y a donc un intérêt à regrouper les séries temporelles de variables environnementales en considérant leurs amplitudes. La même analyse à été faite pour les séries temporelles d'oxygène, de salinité et du renouvellement de l'eau.

Visuellement, les taux de survie des 4 clusters d'oxygène dissous (figure 6.23) ne sont pas corrélés positivement (comme pour la température), aux valeurs d'oxygène dissous;

Concernant la salinité, visuellement, elle impacterait la vitesse de convergence vers le poids final, soit le paramètre b de Gompertz. L'augmentation de la salinité conduirait ainsi à augmenter la vitesse de convergence b vers un poids final d'autant plus faible.

Les clusters précédents concernaient les variables environnementales ou forçantes. Avant de conclure une même analyse a été conduite sur les variables de gestion (renouvellement d'eau et alimentation). Le taux de croissance initial pour chaque cluster (figure 6.25) diminuerait avec un taux de renouvellement plus faible.

Les valeurs des variables temporelles de qualité du milieu, peuvent impacter des élevages de différentes fermes, de manière homogène, en fonction de leurs amplitudes. L'affinement des clusters mono-variés obtenus à partir des séries temporelles des variables de gestions est aussi une source d'informations importantes pour les éleveurs. Elle permet de rechercher des normes de performance à partir des variables temporelles

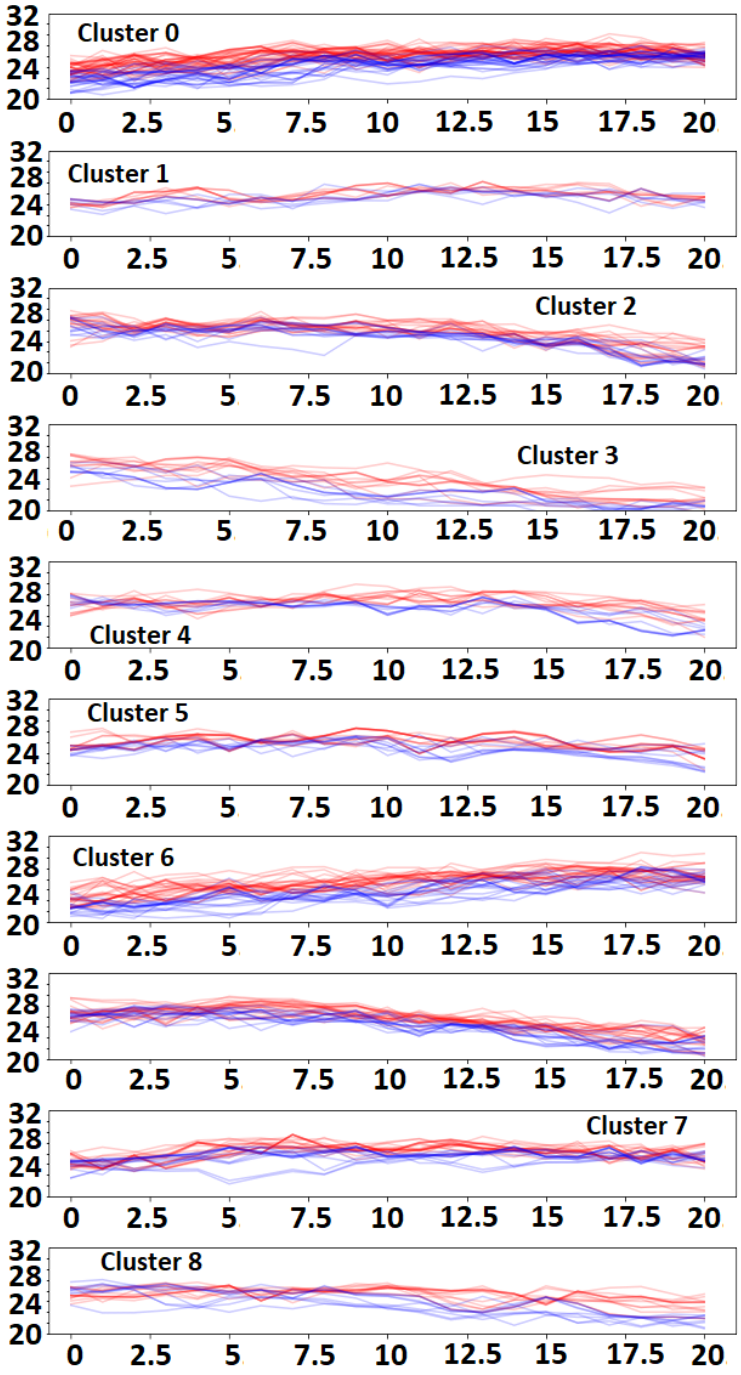

Fig. 6.20 10 Clusters des séries temporelles de température par la méthode *k-shape*

et de manière indépendante, ce qui est le cas de la nouvelle méthode $XmeansTS$ proposée. Par la suite, l'étude montrera que dans le cas des suivis temporelles, il est cohérent de rechercher en premier lieu des normes (des représentants) pour les clusters de variables temporelles qu'elles soient de gestion ou environnementales. L'extraction de représentant, par variables temporelles, va permettre, de créer de nouveau descripteur pour une analyse multi-variée.

Ces nouveaux représentants seront générés, dans le chapitre suivant, par les nou-

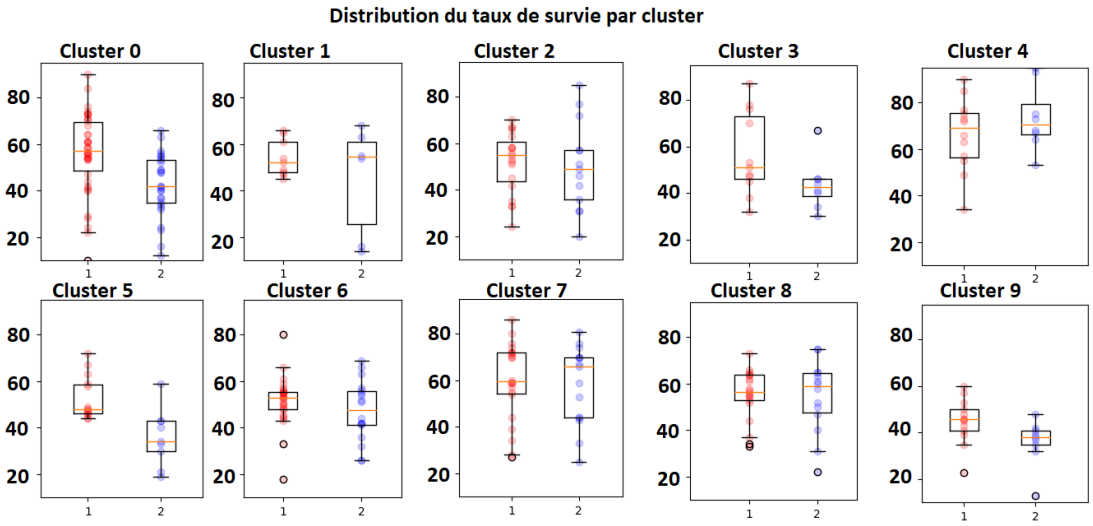

Fig. 6.21 Distribution du taux de survie pour les 10 Clusters de température par la méthode *k-shape*

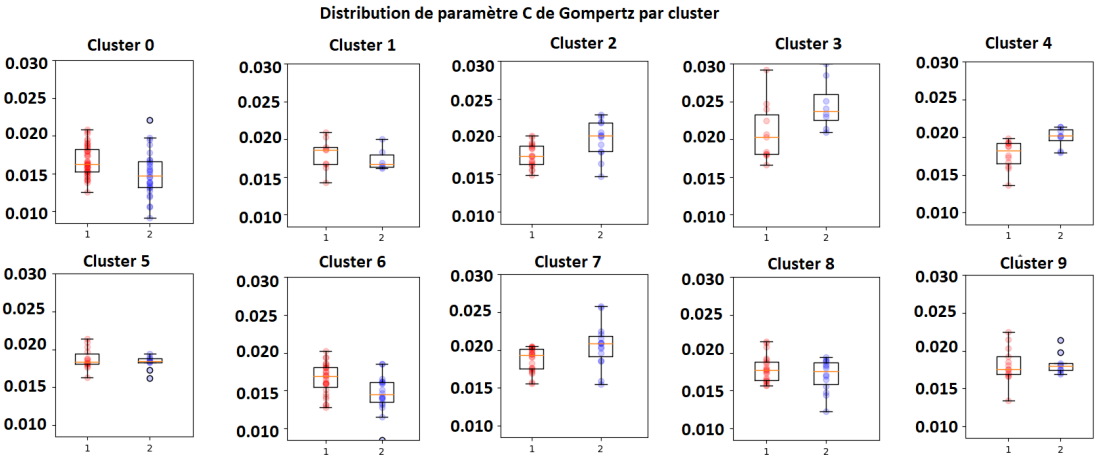

Fig. 6.22 Distribution du taux de croissance initial pour les 10 Clusters de température par la méthode *k-shape*

velles méthodes de clustering de séries temporelles mono-variées et multi-variées, *X-meansTS* et *X-meansMMTS*. Ces clusters seront décris pas les données de performance afin de montrer l'intérêt de ces nouvelles méthodes.

6.3 Conclusion

Nous avons proposé un processus général pour analyser la performance (survie, qualité des produits, etc.) des filières aquacoles en fonction des pratiques d'élevages (mois d'ensemencement, taux de croissance, etc.) et la qualité du milieu (température, etc.). Ce processus vise à intégrer l'ensemble des données complexes (hétérogènes, imprécises, multi-échelles, spatio-temporelles, etc.). Dans ce chapitre, nous avons con-

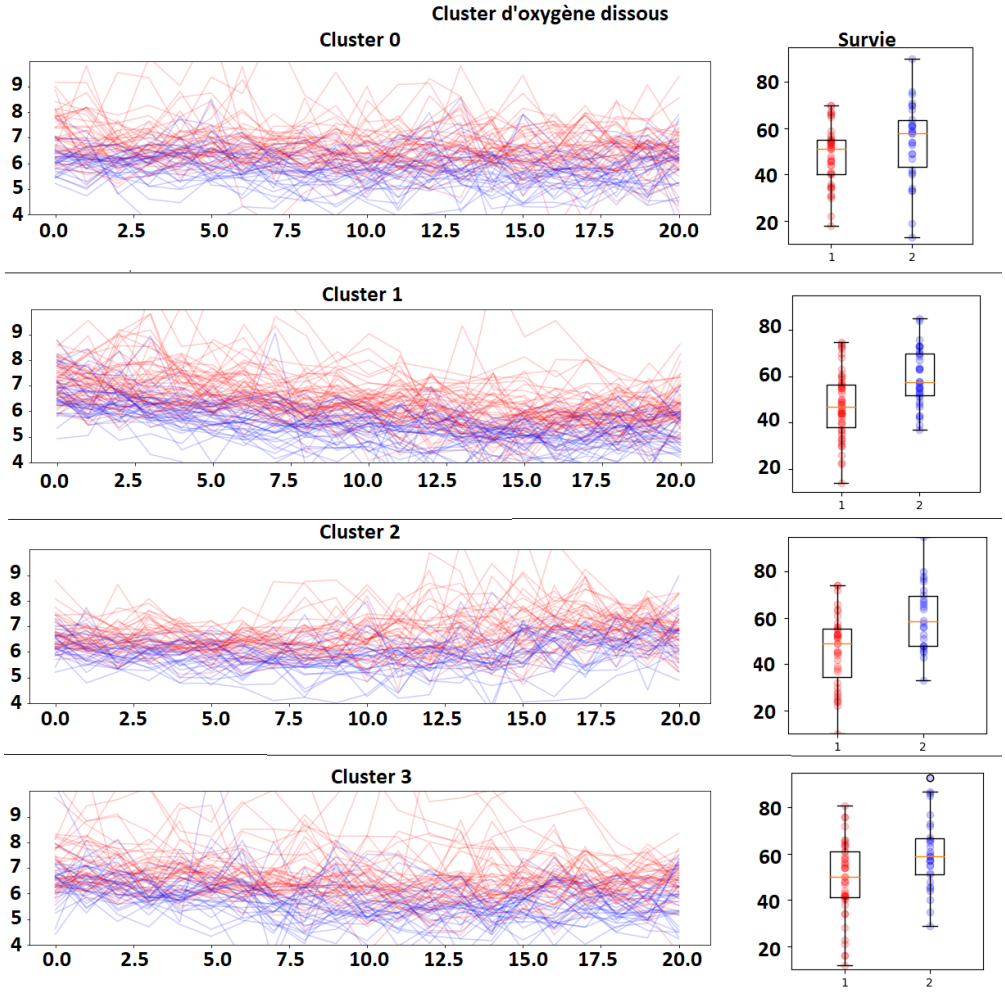

Fig. 6.23 Clustering des séries temporelles d'oxygène par la méthode *k-shape*

tribué à deux étapes de cette méthodologie.

La première étape du processus consiste à analyser des données de croissance en générant de nouveaux descripteurs à partir du modèle de Gompertz appliqué à l'évolution du poids moyen de l'animal. Les descripteurs ont servi d'attributs pour discriminer les élevages. Dans cette étape, les résultats du clustering ont mis en évidence des typologies de croissance qui ont été décrites par diverses données (mois d'ensemencement, performances..). Ces résultats ont montré, en l'occurence, la relation entre des données zootechniques qui décrivent la croissance en début et en fin d'élevage (B,C,PI,...) avec le mois d'ensemencement. Les résultats des différents classifieurs multi-labels sur les mêmes données avec les mêmes descripteurs mais considérant des labels de plusieurs données de performance ont montré que l'on peut mettre en place un modèle prédictif des performances en fonction de la stratégie d'ensemencement appliquée par les éleveurs. Couplée à un modèle économique, cette approche devrait permettre d'optimiser les résultats économiques de cette filière. Ce modèle est évolutif car il peut intégrer des données acquises par les éleveurs chaque année. D'un point de vue méthodologique,

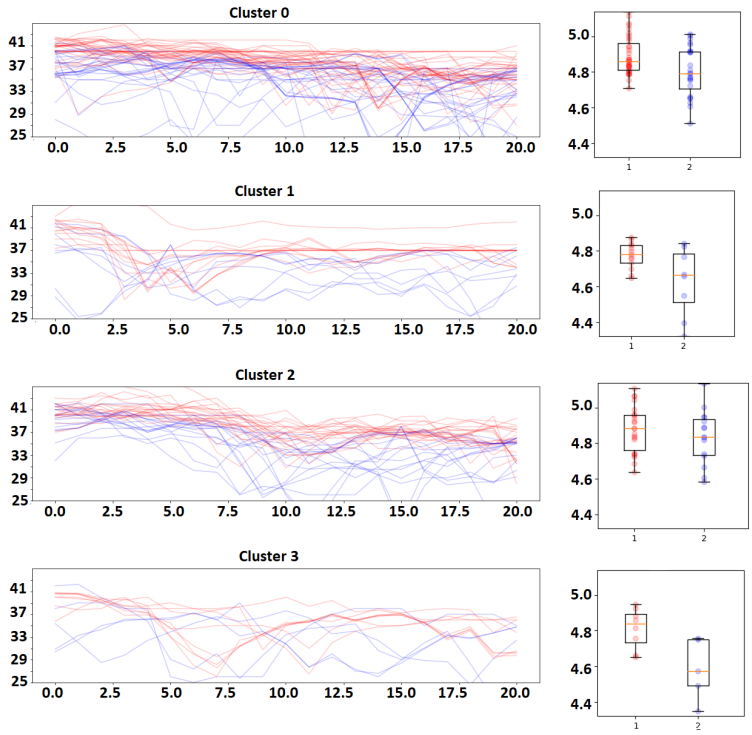

Fig. 6.24 Clustering des séries temporelles de salinité par la méthode *k-shape*

les classifieurs multi-labels qui considèrent les relations possibles entre les labels ont eu de meilleurs performances (precision, recall ...) et notamment *Ensemble classifier chain* qui construit plusieurs classificateurs en chaîne avec un ordre d'étiquettes aléatoire.

La deuxième étape du processus, traitée partiellement dans cette section, a permis de mettre en évidence le lien entre l'évolution de la température durant la totalité de l'élevage et la survie.

Pour compléter le processus d'analyse décrit en introduction, nous proposeront d'analyser le lien entre les données zootechniques, de performances, et de qualité du milieu associée aux séries temporelles de plusieurs paramètres physico-chimiques(salinité, température, oxygène...). L'analyse de ces variables temporelles se fera dans selon des périodes précises, sélectionnées à partir d'indicateur de croissance crée : le jour d'arrivée du point d'inflexion de la courbe de croissance. En effet au début d'élevage la crevette consomme d'avantage pour sa croissance. Au delà d'un certain temps, sa consommation ne sert plus à sa croissance et est rejetée sous forme d'excrément automatiquement. Cette instant peut être déterminé et anticipé par l'arrivée du point d'inflexion par exemple. Le chapitre suivant proposera ainsi une analyse directe des séries temporelles des variables d'environnement et une recherche de liens avec les variables de performances. Cette analyse consiste à rechercher des groupes homogènes de séries temporelles par des méthodes de clustering. Elle se fera, dans un premier temps

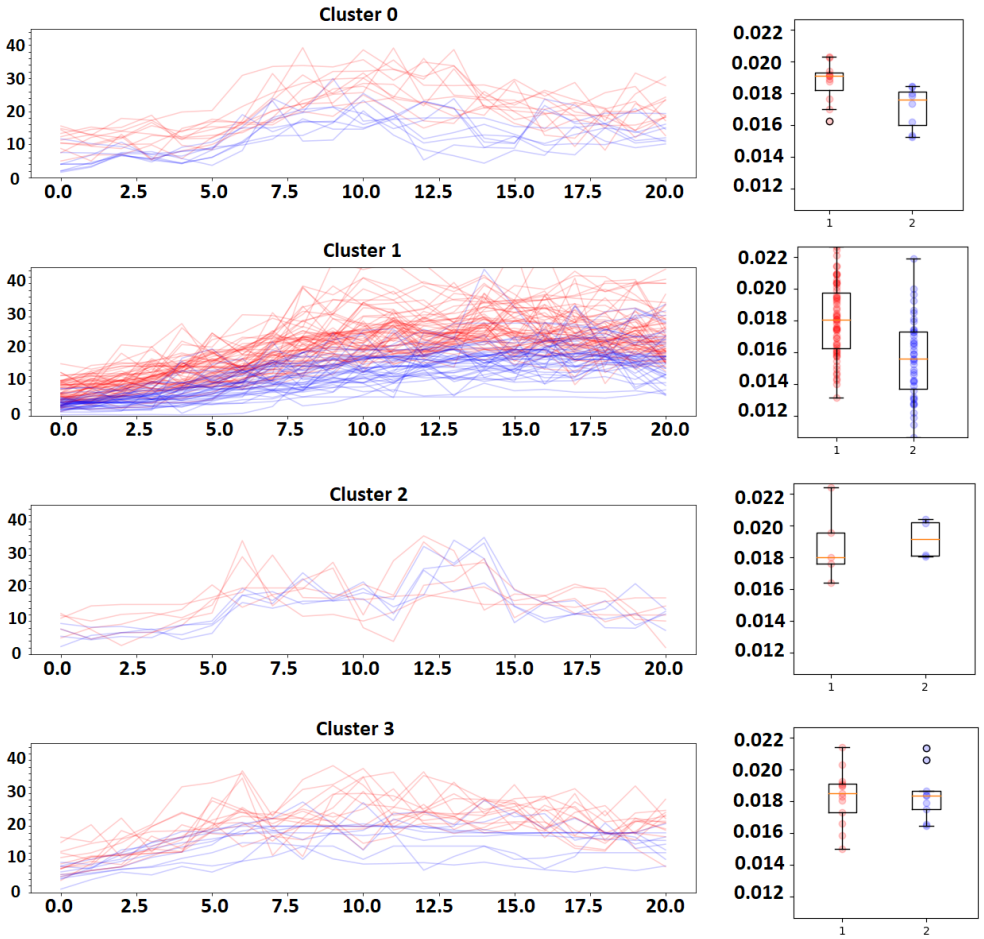

Fig. 6.25 Clustering des séries temporelles du renouvellement de l'eau par la méthode *k-shape*

par variable temporelle, en prenant en compte la variabilité des amplitudes, via notre méthode *X-MeansTS* présentée dans le chapitre **??**; dans un deuxième temps nous proposerons d'appliquer notre deuxième méthode de clustering *X-MeansMMTS* présentée dans le chapitre 3.7 considérant toutes les variables temporelles en même temps

Chapitre 7

Analyse des séries temporelles des variables environnementales et de gestion de la filière aquacole calédonienne.

La nouvelle méthode de clustering de séries temporelles *XmeansTS*, développée et présentée dans le chapitre 3, améliore l'homogénéité des clusters générés par l'intermédiaire d'une méthode (de clustering de séries temporelles) existante. Pour chaque cluster généré, *XmeansTS* affine les clusters, selon l'amplitude des séries. Pour cela trois principaux paramètres sont considérés : la dispersion intra-cluster des distances entre les séries et leur représentant, le nombre de clusters initiaux (générés par la méthode existante), et le nombre minimal d'instance par clusters. Plusieurs tests présentés dans le chapitre contribution, ont mis en avant des résultats satisfaisant sur des ensembles de données libre d'accès en ligne.

Dans le chapitre 2, l'analyse de la qualité du milieu, et de la productivité (croissance et survie), a mit en évidence l'intérêt d'analyser les productions, selon l'évolution, de formes et d'amplitudes des variables de qualité du milieu. La contribution apportée a été de modéliser les relations, sur des jeux de données réelles, entre des clusters de séries temporelles mono-variées et des données statiques (survie en fin d'élevage, taux de croissance initiale vitesse de convergence des crevettes..).

Nous étudierons dans ce chapitre, les données temporelles de la qualité du milieu de la filière, à partir de la méthode de clustering monovarié *XmeansTS*, dans un premier temps. Cette analyse permettra, d'identifier par variable temporelle, des périodes avec un potentiel descriptif de la performance d'élevage. Une analyse muti-variée, sera faite ensuite, par la nouvelle méthode de clustering de séries temporelles multivariées multi-échelles *X-meansMMTS*, à partir de l'interprétation des résultats de l'analyse monovariée.

L'amplitude des séries temporelles environnementales, telle que la température, joue un rôle important sur la physiologie des organismes produits. La complexité à quantifier cet impact est liée pour partie, aux importantes variations de ces variables au cours du temps. Cette section met en évidence, l'intérêt des nouvelles approches face aux données complexes, à générer de nouveaux descripteurs, et à prendre en compte ce type de variation complexe.

Nous présenterons les groupes obtenus à partir de la méthode de *X-MeansTS* avec des résultats interprétés par l'expert en aquaculture. Plusieurs tests ont été effectués et ont été comparés à la méthode *K-shape* en faisant varier les différents paramètres en entrée de la méthode. Nous avons vu dans le chapitre *Contribution* (Chapitre **??**) que la méthode *X-meansTS* générait des clusters de qualité, sur des données disponibles en ligne sur le site *UCR* ([7, 21]). Néanmoins les données temporelles générées dans les filières aquacoles sont complexes (i.e les séries varient fortement). Nous avons proposé une amélioration de la méthode *X-MeansTS*, en suivant le même principe de la méthode (une variante que l'on détaillera dans la section 7.1.1 suivante), dans laquelle, on discrétise la distribution des distances $Dist(C)$ par cluster. Nous verrons que cette variante fournie de bons résultats sur les séries temporelles des données d'aquaculture.

7.1 Optimisation de la méthode *X-MeansTS* par la discrétisation des distances intra-clusters

7.1.1 Principe de l'approche discrétisée

Avant de détailler le principe de l'approche discrétisée, rappelons le principe de l'approche non discrétisée : dans l'approche non discrétisée, les séries sont intégrées successivement dans le calcul de la mesure de dispersion d_p tant que la mesure est inférieure au seuil de dispersion. Cette intégration se fait dans l'ordre croissant des distances entre les séries d'un cluster et le représentant. En appliquant *X-MeansTS*, sur ces données réelles, on remarquera, en raison des fortes variations de ces données, que dans un cluster, deux séries successivement proches de leur représentant, peuvent avoir une distance *DTW*, avec celui-ci, qui reste élevée. Cette contrainte ne permet pas de générer, avec une méthode telle que *K-Shape*, ou *K-MeansDTW*, des clusters hétérogènes, à condition de paramétrer un nombre de cluster souhaité k très grand. Dans ce dernier cas le nombre d'individus par cluster, peut être faible. Lorsque les séries varient fortement, les clusters générés par *X-meansTS* avec un seuil de dispersion s_d très faible, peuvent contenir des séries temporelles avec des tendances observables. Or, avec ce seuil (faible), ces clusters peuvent avoir très peu d'individus voire, peuvent contenir une seule série. L'avantage avec l'approche discrétisée est de sélectionner un nombre d'individus avec un seuil faible, tout en considérant la dispersion des distances intra-clusters.

Par cluster, dans l'approche par discrétisation, les distances entre les séries et leurs représentant sont normalisées. Un nombre commun de séries, les plus proches du représentant, est déterminé. A partir de ces séries l'étendue entre la distance minimale et maximale (normalisées), servira à déterminer l'intervalle de référence pour discrétiser les distances de toutes les séries d'un cluster. Par exemple si l'on fixe à 3 le nombre minimale de séries par cluster, les distances des séries du cluster sont d'abord normalisées

et la première classe contiendra les 3 premières séries les plus proches du représentant selon la mesure de distance utilisée dans *X-meansTS*. La classe de ces séries sera la distance moyenne de ces 3 premières séries au représentant (la moyenne est calculée par rapport à la distance discrétisée des séries). C'est cette valeur moyenne qui sera comparé au seuil de dispersion afin de déterminer si ces séries sont sélectionnées ou rejetées du clusters. Pour obtenir les classes liés aux séries suivantes du clusters : à partir de ces 3 premières séries, un intervalle de référence est calculé. La borne inférieur de cet intervalle de référence correspondra à la distance de la série la plus proche du représentant, et la borne supérieur à plus éloignées. La classe suivante contiendra les séries dont la distance, avec le représentant, sera comprise entre la valeur de la borne supérieur de l'intervalle de référence et 2 fois cette même valeur (de la borne supérieur de l'intervalle de référence). La classe des séries, comprises dans cet nouvel intervalle, sera la valeur moyenne des séries au représentant. Comme exemple, la figure 7.1 schématise l'approche discrétisée de *X-meansTS* avec avec nb_min_inst = 3.

Pour discrétiser nous procédons aux étapes suivantes, sur chaque cluster générés initialement par une méthode existante (telle que *K-meansDTW* dans *X-meansTS*. Soit C un cluster initial:

- Étape 1) Création de $N(DTW(C)) = \{v_1, v_2, \ldots, v_m\}$ ou $v_i = \frac{v_i}{max(DTW(C)) - min(DTW(C))}$ avec $max(DTW(C))$ et $min(DTW(C))$ respectivement les valeurs maximal et minimal des mesures de $DTW(C)$: Les valeurs de DTW de chaque cluster "initiaux" sont normalisés.

- Étape 2) Ordonnancement de $N(DTW(C))$: $O_N(DTW(C)) = \{v_1, v_2, \ldots, v_m\}$ avec $\{v_i, v_j \in N(DTW(C)) \mid i < j \ , \ v_i < v_j\}$: Les valeurs des séries normalisées de chaque cluster sont ordonnées selon l'ordre croissant.

- Étape 3) A partir de la série ordonnée et normalisée $O_N(DTW(C))$ on sélectionne dans l'ordre de la série, un intervalle constitué d'un nombre de valeur égale au nombre minimal d'instance $min(c_1^l)$. La borne inférieur $b_{inf}(C)$ et la borne supérieur $b_{sup}(C)$ de cette intervalle servira d'écart entre plusieurs intervalles successives dans $O_N(DTW(C))$. Ces intervalles seront les classes (cf étape 4).

- Étape 4) Définition des instances par classes : On obtient $cl_j(C)$ la classe j qui contient les valeurs $V \in O_N(DTW(C))$ tel que $V = \{v_i, v_{i+1}, \ldots, v_{i+h}\} \subset [j * b_{inf}, j * b_{sup}]$.

- Étape 5) dans le calcul de la mesure de dispersion de $d_p(C))$ les distances seront remplacées par la moyenne des individus de leurs classes respectives.

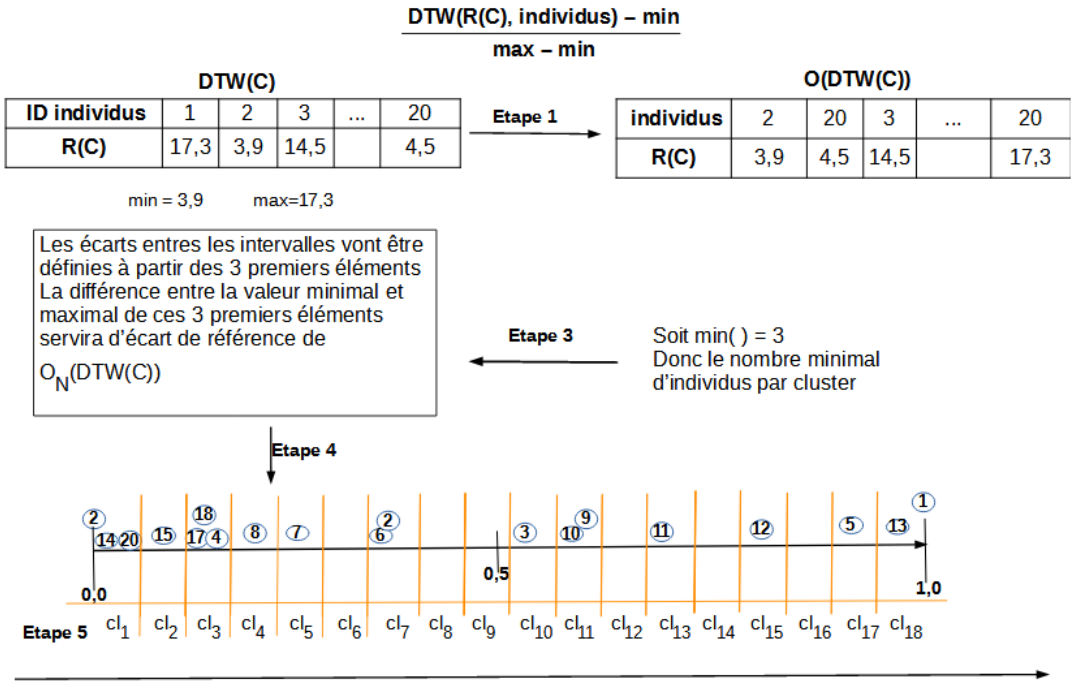

Fig. 7.1 Optimisation de l'approche X-meansTS par discrétisation

7.2 Résultats de l'Application de la méthode *X-MeansTS* sur les données temporelles de la filière aquacole

7.2.1 Description des clusters et comparaison avec des méthodes existantes sur les données de la filière aquacole calédonienne

Des clusters ont été générés par variable de qualité du milieu d'élevage (variables temporelles environnementales et de gestion telles que la température, l'oxygène dissous, le renouvellement d'eau ...), par l'approche discrétisée de la méthode *X-meansTS*.

7.2.2 Comparaison visuelle des clusters générés par *X-meansTS* et *K-Shape*

La méthode *X-MeansTS* a la particularité de générer des représentants fiables, et notamment sur des séries avec d'importantes variations. Les clusters générés par *X-MeansTS* a été comparé visuellement au clusters générés par la méthode *K-Shape*; La méthode *K-Shape* a été choisie car, elle obtient d'excellentes performances pour le clustering de séries temporelles sur des données de *Benchmark* (cf. chapitre **??**). Les figures 7.2 et 7.3 montrent, par exemple, pour la salinité et l'apport en aliment, que cette approche, permet de générer des clusters avec un nombre d'individus assez important, et avec des tendances homogènes. L'approche *K-shape* peine à déterminer ce type de clusters. Il apparaît que les représentants des clusters sont plus proches de leurs individus pour la méthode proposée *X-MeansTS*, que les représentants des clusters obtenus par *K-Shape*.

Les clusters contiennent donc des séries avec des tendances plus homogènes.

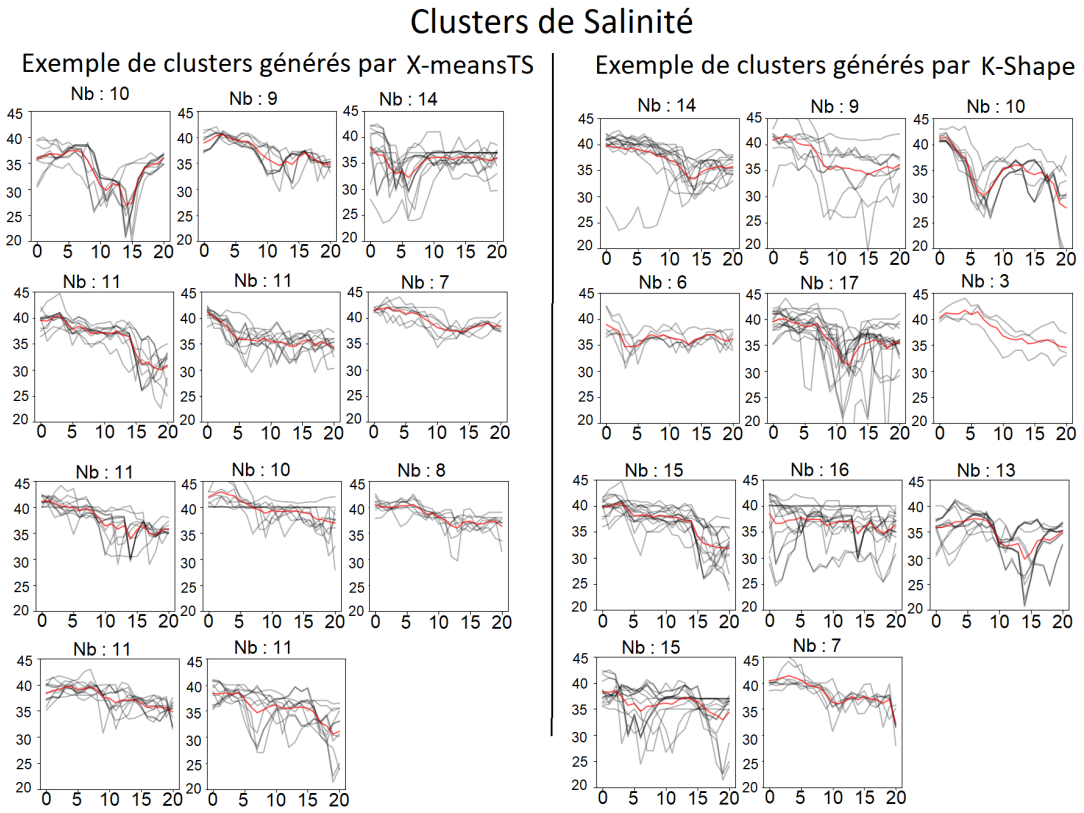

Fig. 7.2 Clusters de salinité obtenus par la méthode *X-meansTS* à gauche et la méthode *K-Shape* à droite

7.2.3 Méthode de comparaison statistique des clusters générés par *X-meansTS* et *K-Shape*

Les clusters obtenus par *X-meansTS* ont été interprété par un expert du domaine. Une interprétation des résultats, se fera, par cluster, au travers d'une analyse des distributions des variables explicatives de performance des élevages via des statistiques descriptives (BoxPlot). Les hypothèses générées dans cette analyse seront validées par des tests statistiques (ANOVA). Nous testerons la différence des clusters en les comparant par pair de clusters en prenant en compte les variables de qualité de production. Le test statistique, rejettera ou non l'hypothèse selon laquelle les moyennes, de deux distributions, d'une même variable de performance seraient proches (i.e une p-valeur < 0.05).

Les pairs de clusters générés par *X-meansTS*, la distribution des variables de performance est significativement très différentes seront retenus pour l'interprétation. Elles seront comparés aux pairs de clusters, générés par la méthode *K-Shape* qui présentent également des différences significatives des distributions des variables de performances qui les décrivent.

.

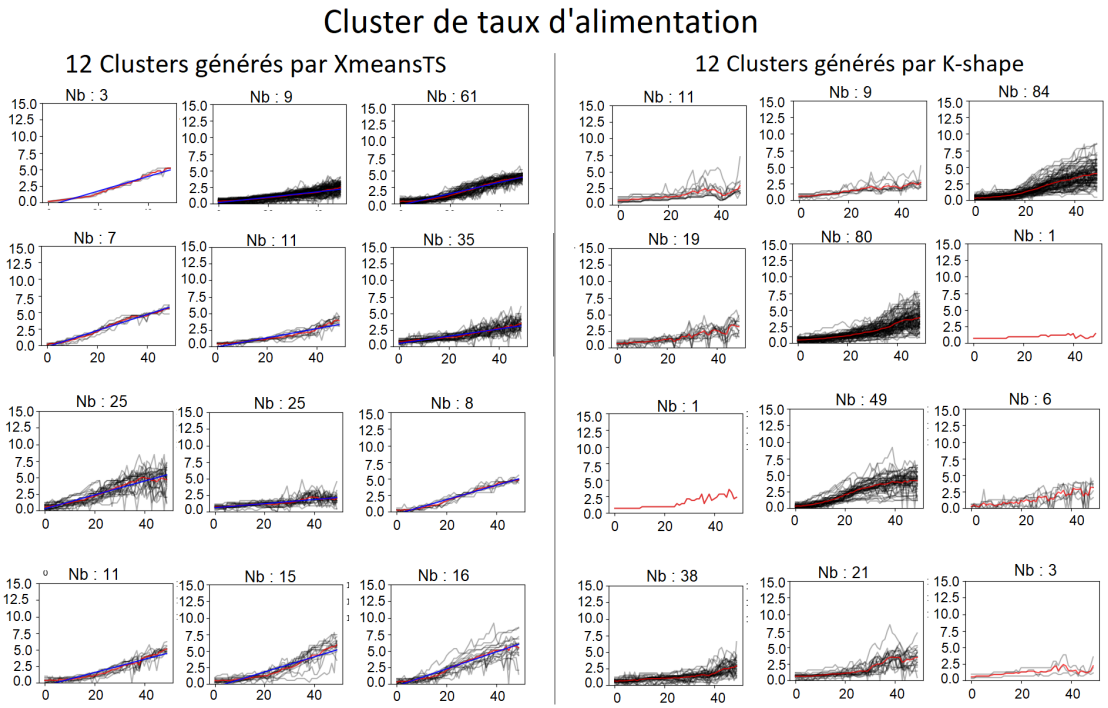

Fig. 7.3 Comparaison des clusters pour l'apport en aliment journalier obtenus par la méthode *X-meansTS* à gauche et la méthode *K-Shape* à droite

7.2.4 Choix des paramètres en entrée de la méthode *X-MeansTS*

Pour obtenir les clusters de qualité du milieu par la méthode *X-meansTS*, différents paramètres en entrée on été testés. Pour chaque variable temporelle de qualité d'eau, différents résultats de clustering seront présentés en faisant varier les différents paramètres en entrée de la méthode *X-meansTS* (seuil de dispersion d_p, nombre minimum de clusters nb_min_clust, etc.).

Le paramètre nb_min_clust (nombre de clusters initial) sera fixé à 5 (qui sera ensuite modifié selon la nouvelle mesure de dispersion). En effet, le chapitre précédent a mit en évidence 5 typologies de croissance, au minimum. Il a aussi mis en avant un lien potentiel entre les variables de qualité du milieu, le taux de croissance initial et la vitesse de convergence vers le poids final. Ces liens, pour rappel, ont été perçus en redécoupant les clusters de séries temporelles de température (voir section **??**), en tenant compte des aires sous les courbes (des séries).

Le nombre minimum d'instances par cluster nb_min_inst sera fixé à 10. Les clusters seront générés en utilisant des seuils de dispersions dits minimaux, moyens et maximaux. Afin d'obtenir ces seuils automatiquement, par variable, les séries sont regroupées à plusieurs reprises en différents nombres de clusters, par les méthodes existantes *K-shape* ou *K-meansDTW*. Par cluster, la mesure de dispersion est appliquée aux distances entre les individus et leur représentant, pour obtenir une gamme de seuils de dispersion. A partir de ces seuils, les seuils de dispersion minimaux, moyens et max-

imaux sont obtenus pour être utilisés (indépendamment) en tant que paramètre d'entrée de la méthode *X-meansTS*.
Les valeurs de ces seuils, seront propres à chaque variable temporelle. La figure 7.4 présente les seuils moyen, minimal et maximal en fonction des variables de température, d'oxygène, de renouvellement, de salinité et d'alimentation. Pour rappel , ces seuils sont obtenus à partir de clusters générés dans la méthode par la méthode *K-shape*, et auxquels le calcul de la mesure dispersion *Disp()* est appliqué pour obtenir une liste de seuil. A partir de cette liste les valeurs de seuil minimal, et moyen sont obtenus. Dans l'approche discrétisée, les seuils doivent être modifiés car les distances des séries sont modifiés (discrétisées). Pour les tests, les seuils conservés (minimal et moyen) sont divisés par le nombre d'instance minimal nb_min_inst (considérée pour obtenir les classes auxquelles appartiennent les séries).

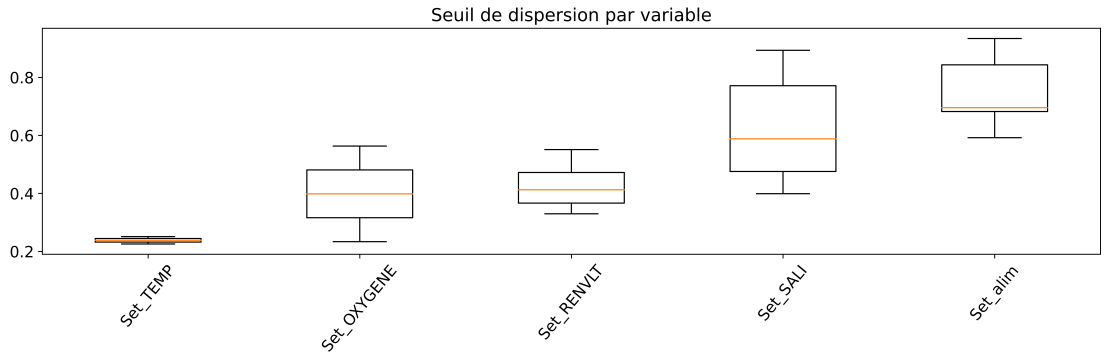

Fig. 7.4 Distribution des seuils en fonction des variables

Remarque : Des comparaisons entre les performances des méthodes *X-meansTS* et *K-Shape* ont donc été réalisées sur des nombres identiques de clusters générés par ces deux approches. Pour cela le nombre de clusters (finaux) générés par *X-meansTS* a servi de paramètre *k*, en entrée de la méthode *K-Shape* comme nombre de clusters souhaité.

7.2.4.1 Choix de la période d'élevage avec un potentiel descriptif de performance d'élevage.

L'hypothèse que nous émettons ici est que les performances d'élevage seraient significativement plus impactées par certaines périodes d'élevage. Le chapitre précédent 6 a mis en évidence un lien potentiel entre la température et les données de productivité (paramètres de croissance, survie..). Le point d'inflexion de la courbe de croissance, lié à la vitesse de croissance initiale, et qui a été étudiée dans le chapitre 6, permettra de déterminer l'intervalle de la période d'élevage à considérer pour l'analyse des séries temporelles de qualité du milieu (température, salinité,..) et de qualité de d'élevage. A titre d'exemple, les animaux juste après l'ensemencement pourrait être

particulièrement sensibles à des paramètres environnementaux comme la température. La période ciblée est le début de l'élevage. L'étude sera faite sur une période, commune à l'ensemble des élevages, avant le point d'inflexion de la courbe de croissance (modèle de Gompertz). Les clusters de croissance du chapitre précédent contenant principalement les élevages ensemencés en début d'année, montraient des courbes de croissance dont le point d'inflexion se présentait assez tôt. Pour ces clusters, la moyenne (en jour) d'arrivée de ce point, était de 50 jours. En moyenne l'arrivée la plus tardive, du point d'inflexion était de 120 jours. Nous considérerons par la suite, une analyse sur les 50 premiers jours pour les variables temporelles avec une fréquence journalière, et 10 semaines pour les variables avec une fréquence hebdomadaire, en tant que période avant le point d'inflexion de la courbe de croissance.

7.2.5 Interprétation des clusters générés par *X-meansTS*, sur les variables environnementales

La méthode *XmeansTS* (méthode de clustering monovariée) sera appliquée aux variables de qualité du milieu (une par une), notons que l'interprétation des clusters, déterminera s'il existe une corrélation, entre chaque variable de qualité du milieu et de productivité. Une analyse indépendante de variable de qualité (clustering mono-varié) mettra en évidence des corrélations spécifiques entre chaque variable de qualité du milieu et la productivité.

Analyse de l'impact de la température sur la croissance avec les clusters générés par *X-meansTS* : Cette section présente une première analyse avec les courbes de moyennes de la température hebdomadaire et une seconde avec les données journalières par la nouvelle méthode *X-meansTS*. Les représentants de clusters seront décrits par des données statistiques (ex. min, max, moyenne) et ils seront associés aux données de performances (croissance, survie).
La figure 7.5 affiche les individus et leur représentant (trait rouge), pour les 10 clusters de température hebdomadaire avec un seuil de dispersion minimal à gauche et moyen à droite.
Nous nous intéresserons davantage aux clusters générés avec le seuil moyen. L'interprétation fournit par les résultats du seuil minimal, est généralisable aux clusters obtenus par le seuil moyen. Avec ce dernier, il y a davantage d'individus par cluster. Pour ce seuil, la figure 7.6 présente les *p-valeurs* calculés par pair de cluster et en fonction des variables de performance suivantes :

- le taux de croissance initiale C,
- la vitesse de convergence vers le poids final B,

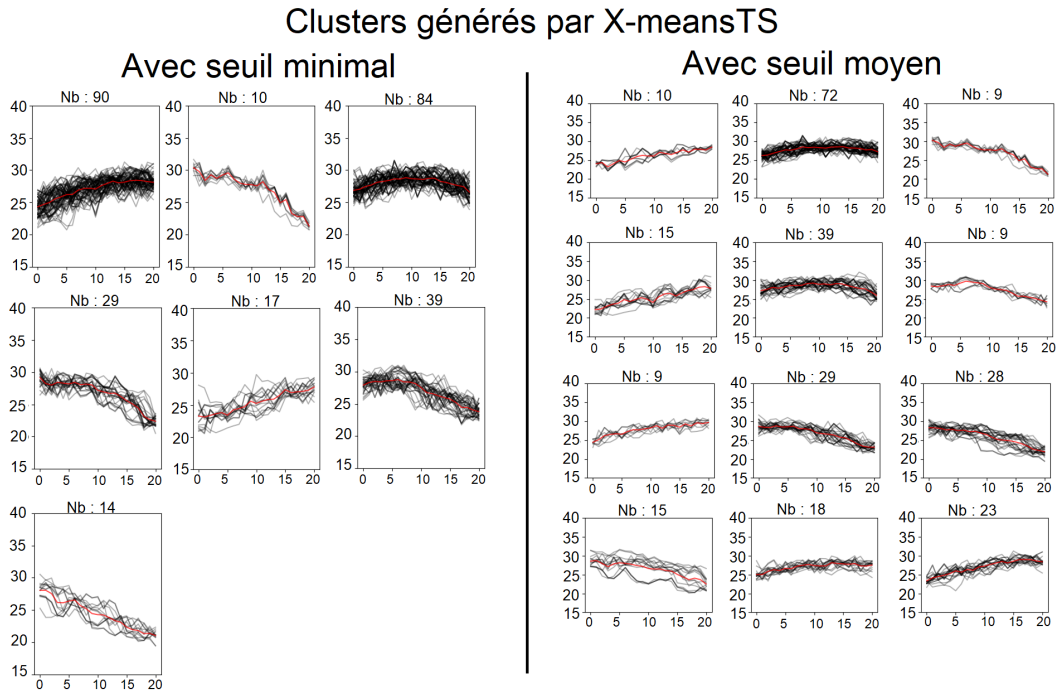

Fig. 7.5 Clusters de température obtenus par la méthode *X-meansTS*

- la survie,
- et le mois d'ensemencement, qui présente les p-valeurs les plus faibles;

Ces tests confirment que le mois d'ensemencement est la variable explicative prépondérante, de l'évolution de la température de l'eau.

Nous verrons ensuite que les distributions sont visiblement distantes, pour toutes les pairs de clusters (obtenus par *X-meansTS*), dont la *p-valeur* est inférieur à 0.05. Par exemple la figure 7.7 montre les pairs de clusters, 3/1 et 8/4 avec les p-valeurs inférieures à 0.05.

Le tableau 7.1 présente les valeurs minimales, maximales et moyennes des représentants de certains clusters de la figure 7.8, et ces mêmes statistiques pour les variables de performances (survie, croissance). Les écart-types des survies sont élevés, confirmant que ce niveau de résolutions ne permet pas d'émettre une interprétation objective. On remarque aussi, que lorsque la température hebdomadaire est plus élevée, et notamment avant la $10^{ème}$ semaine, le taux de croissance initiale (paramètre C de Gompertz) serait plus élevé et la vitesse de convergence la plus faible (figure 7.7). Ce constat a été le même pour toutes les pairs de clusters dont la *p-valeur* calculée sur les variables de Gompertz était inférieur à 0.05 (Cf exemple de pairs de clusters en Annexe A).

Afin de confirmer ou non ces affirmations, une analyse réalisée en prenant en compte non plus la moyenne hebdomadaire de la température mais les valeurs journalières. La période considérée concerne les valeurs avant le point d'inflexion de la courbe de croissance.

	Température hebdomadaire (°)				Taux de survie			
Clust	*min*	*moy*	*max*	*écart-type*	*min*	*moy*	max	*écart-type*
1	26.21	27.83	28.57	0.69	10.0	55.32	84.0	12.6
3	22.2	25.4	28.28	1.83	18.0	44.07	69.0	15.48
4	25.98	28.24	29.22	0.83	24.0	58.36	90.0	13.21
8	21.92	25.86	28.31	2.09	13.0	46.57	81.0	17.18

Table 7.1 Statistiques des représentants de clusters de température hebdomadaire et moyenne de survie

D'après la figure 7.8, comme on devait s'y attendre, plus la température de l'eau est élevée en début d'élevage, et plus la vitesse de croissance initiale parait élevée (i. paramètre *C* de Gompertz élevé). Mais on note aussi une baisse de la vitesse de convergence vers le poids final. Visuellement, et de fait, les calibres ont tendances à être plus petit.

Ces types de relations ayant été relevées dans la littérature, il s'agit, dans cette thèse de préciser les distributions des variables explicatives, avant le point d'inflexion de la croissance, en fonction des clusters et des données statistiques de leur représentant (valeur minimale, maximale, moyenne). Nous préciserons tout de même, par pair de cluster, que les distributions significatives, des paramètres de Gompertz, sont plus distantes, pour les clusters de températures prises avant l'inflexion de la courbe de croissance (Cf exemple de pairs de clusters en Annexe B);

D'après les représentants des séries prises avant le point d'inflexion de la courbe de croissance, on observe qu'une température de l'eau plus importante, n'impliquerait pas obligatoirement un taux de survie plus élevé. Il apparaît que le taux de survie serait supérieur en moyenne à 50%, lorsque la température se situerait entre 26° et 27° dans les 50 premiers jours. Dans cet intervalle de temps, une décroissance de la température vers des valeurs en deçà de 26° impacterait le taux de survie; Au contraire une température journalière qui croit, ou qui est stable dans cet intervalle de valeur, et durant les 50 premiers jours, favoriserait un meilleur taux de survie. Néanmoins les écarts types des taux de survie par clusters sont proches de ceux des clusters des séries de température hebdomadaire, prises sur l'ensemble de l'élevage.

Or ces "normes" de croissance, dépendent évidemment d'autres facteurs environnementaux qui, nous le verrons par la suite, sont dépendant de facteurs de gestion (taux et fréquences, de nourrissage et de renouvellement d'eau). Et un modèle descriptif multi-échelle, impliquant l'analyse, de la température, à différents niveaux de résolutions et à différentes périodes, déterminera si, entre ces niveaux et ces périodes, l'évolution de température de l'eau, par élevage, est homogène ou hétérogène, et si ces variations peuvent constituer un facteur explicatif pour le taux de survie finale.

Concernant spécifiquement la qualité du produit, les *p-valeurs* calculés sur les distri-

butions de ces défauts (par pairs de clusters), ne sont pas significatives. Ils ne permettent pas d'établir un éventuel effet.

Clusters de la variable temporelle de salinité par la méthode *X-MeansTS* La variation de la salinité au cours du temps est très importante. Obtenir des groupes très hétérogènes entre cluster, avec les méthodes de clustering de séries temporelles existantes, est une tache complexe. En comparaison avec les résultats précédents liés aux clusters de température, les résultats du clustering des séries de salinité sur l'ensemble de la période d'élevage, nous pousse à analyser ces séries avant le point d'inflexion de la courbe de croissance. En effet selon ces clusters, la croissance pourrait être corrélée à la salinité dans les 10 premières semaines; Les résultats du clustering des séries avant l'inflexion de la croissance, (figure 7.9) montre que, plus la salinité serait élevée plus la croissance initiale (C) ralentirait. Ce résultat est cohérent avec la bibliographie qui montre que les animaux dépensent plus d'énergie lorsque la salinité augmente [101]. La relation entre la vitesse de croissance initiale (C) et les calibres n'est pas observable comme pour la température. En effet, l'écart des moyennes des distributions de cette variable, est faible entre clusters. La salinité impacterait peu les calibres.

Clusters de l'oxygène dissous, par la méthode *X-MeansTS* par discrétisation La variation de l'oxygène dissous dans les bassins d'élevage au cours du temps est très importante. Pour cette raison, il est complexe d'obtenir différents groupes hétérogènes avec les méthodes de clustering de séries temporelles existantes (voir en Annexe C 'comparaison avec *K-Shape*'). Quelques effets sont observables avec la méthode *X-meansTS*, entre les clusters d'oxygène dissous, et les paramètres de croissance et de survie.

7.2.6 Interprétation des clusters générés par X-meansTS, à partir des variables de gestion

Pour rappel les résultats qui seront présentés, sont des pairs de clusters générés par *X-meansTS*, qui ont par variable explicative, des distributions significatives.

Clusters liés à la variable de renouvellement de l'eau par la méthode *X-MeansTS* selon l'approche par discrétisation : La figure 7.10 montre les distributions des paramètres zootechniques, en fonction des clusters 3, 2 et 8 du renouvellement d'eau, obtenus avec un seuil minimal de dispersion. Entre les pairs 3/2 et 3/8, les distributions de chaque paramètre de croissance B et C, sont significatives; Cependant, la figure 7.10 ne permet pas de préciser si le taux de renouvellement pris sur l'ensemble de l'élevage, est corrélé positivement ou négativement à la croissance initiale, ou à la vitesse de convergence vers le poids final. C'est à dire, si un renouvellement plus régulier favoriserait une

Cluster	min	moy	max	coefficient de pente	ordonnée à l'origine
1	3.45	10.38	16.93	1.447	3.145
6	4.97	13.96	25.81	2.001	3.951
7	10.13	16.27	23.48	1.471	8.912

Table 7.2 Valeur moyenne des représentants de clusters de renouvellement hebdomadaire d'eau, avant l'inflexion de la courbe de croissance

	Survie				**b**			
Clust	*min*	*moy*	*max*	*écart-type*	*min*	*moy*	max	*écart-type*
1	28.0	48.6	53.0	11.35	4.72	4.86	5.06	0.12
6	42.0	53.82	75.0	10.53	4.64	4.83	5.02	0.1
7	47.0	63.8	81.0	12.24	4.32	4.7	4.92	0.17

Table 7.3 Données de productivité liées aux clusters de renouvellement d'eau, avant l'inflexion de la courbe de croissance

augmentation ou une réduction des paramètres.

En conséquence, un clustering avant l'inflexion de la croissance (i.e sur les 10 premières semaines) permettra d'affiner notre compréhension du lien entre le taux renouvellement et la croissance initiale.

Impact du renouvellement d'eau, les 10 premières semaines d'élevage, sur la croissance initiale et la survie : La figure 7.11 présente les clusters de séries de renouvellement d'eau prises sur les 10 premières semaines d'élevage; Les données statistiques (cf moyenne, minimale, maximale tableau 7.2), des représentants qui sont affichés en couleur rouge, et leur modèle de régression linéaire en bleu serviront, par la suite, à comparer ces clusters. Les modèles de régression des représentants, seront décrits par la pente et l'ordonnée à l'origine (tableau 7.2).

D'après la figure 7.12, et le tableau 7.2, la vitesse de croissance initiale, parait plus lente lorsque le renouvellement d'eau en début d'élevage (avant l'inflexion de la croissance) est moins important. Le tableau 7.3, présente les statistiques des données de qualité par clusters (7, 6 et 1 de la figure 7.11). Les écarts types des distributions des séries, sont (en considérant tous les clusters), beaucoup plus faibles que les écart types des taux de survies des clusters de température. La fréquence de renouvellement d'eau impacterait le taux final de survie, plus significativement que sa température.

Les résultats du clustering des séries de renouvellement prises sur les 10 premières semaines, approuveraient que le taux de croissance initiale de la crevette serait corrélé positivement au renouvellement d'eau, avant l'inflexion de la courbe de croissance.

Clusters de l'apport en aliment au cours du temps, par la méthode *X-MeansTS* : L'analyse des séries, de taux journalier d'alimentation, prises sur 150 jours d'élevage,

Variable temporelle	C	B	survie	calibre
Variable environnementale				
température 50 jours	P	N	P	P
oxygène dissous	X	X	X	X
salinité 10 sem	N	P	X	X
Variable de gestion				
alimentation 50 jour	X	P	P	X
renouvellement 10 sem	P	N	P	X

Table 7.4 Type de corrélation entre séries temporelles de qualité du milieu, prises à différentes périodes et les données de qualité de production

a montré le lien naturel entre ce taux avec la croissance et la survie. Ces variables de performance sont positivement corrélées à la quantité d'aliment apportée dans le temps; En effet, la survie et la vitesse de croissance initiale, augmentent avec l'apport en aliment;

7.2.7 Synthèse sur l'analyse de la qualité du milieu par X-meansTS

La comparaison de la méthode a permis de déterminer, entre ces deux méthodes, le nombre de pair de cluster dont les distributions, de différentes variables de qualité, sont significatives. Les clusters affichés ont été ceux pour lesquels les distributions des paramètres extraites du modèle de Gompertz, C et b, et de la survie, ont donc des p-valeurs inférieures à 0.05. Notons que la méthode *K-Shape* a généré très peu de pairs de clusters respectant cette contrainte. La nouvelle méthode proposée à la particularité de créer des représentants proches des tendances des évolutions réelles de la qualité du milieu des élevages. Les résultats de l'analyse mono-variée ont mis en évidence, certaines corrélations entre les variables de qualité du milieu, et les données de performance. Le tableau 7.4 les affichent lorsque la corrélation est positive ou négative. On remarque l'intérêt d'analyser différentes périodes d'élevage, qui pour ces variables temporelles, permettent de préciser si la corrélation est positive ou non. C'est en effet le cas de la température, de la salinité, et du renouvellement en eau.

Après que la méthode *X-meansTS* ait affiné les clusters de séries temporelles en fonction de l'amplitude, nous avons pu créer de nouveaux descripteurs fiables, représentatifs de la qualité du milieu des élevages. Les données statistiques de ces représentants de clusters, et de leurs modèles linéraires, considérés avant l'inflexion de la courbe de croissance, offrent la possibilité de créer des outils d'aide à la gestion des productions qui tiennent compte de différentes périodes. Ces périodes ont été déterminées dans le chapitre précédent et selon les paramètres de croissance initiale et la vitesse de convergence vers le poids final.

Afin d'approfondir cette compréhension une analyse multi-variée, devait être réalisée.

7.3 Clustering de séries temporelles multi-variées et multi-échelles sur la qualité du milieu d'élevage aquacole

Dans cette section, les résultats présentés, sont issus d'une analyse multi-variée multi-échelle (MMTS) des séries temporelles qui ont été traitées indépendemment dans la section précédente. La nouvelle méthode de clustering de séries temporelles multi-variées et multi-échelles, détaillée dans le chapitre "Contribution", sera appliquée aux variables temporelles disponibles dans la base de données de la filière crevetticole Calédonienne; Nous nous intéresserons aux variables temporelles définissant la qualité de l'eau. Pour rappel, dans l'approche multi-variée, de nouveaux descripteurs seront crées à partir de *X-meansTS*, indépendamment sur chaque variable temporelle. Ces nouveaux attributs, comme expliqué dans le chapitre "Contribution", sont les représentants, des clusters obtenus, par variable. Ils serviront d'attributs pour la nouvelle méthode de clustering multi-variées. Les individus (les élevages), seront donc décrits par des variables communes de qualité du milieu. Les valeurs de ces attributs seront des données statiques qui correspondent aux distances entre les variables temporelles communes aux élevages, et les représentants (de chacune des variables). Étant donnée que les valeurs entres les variables ne sont pas enregistrées à la même fréquence, le clustering multivarié est donc également multi-échelle. Néanmoins une analyse multi-variée faisant intervenir tous les élevages intégrera peu d'instances. Pour rappel, peu d'élevages ont assurés un suivi complet de toutes les variables décrites dans le chapitre 4 '*Description des données*'

L'intérêt d'une analyse multi-échelle pour des données environnementales : De nouvelles connaissances extraites des données temporelles de la filière crevetticole calédonienne, à partir de la précédente analyse mono-variée, ont permis d'évaluer quantitativement l'impact d'une variable de qualité du milieu sur une variable de qualité de production, et à des périodes pertinentes durant l'élevage. L'analyse multi-variée, vise à apporter plus de précision des effets conjugués de ces variables, et à différents niveaux de résolution temporelle, sur la qualité de production à partir des périodes analysées dans la section précédente (température en début d'élevage, 10 premières semaines de salinité ...). Notons que l'analyse multi-échelle peut extraire davantage d'informations pertinentes. Considérons comme exemple, un clustering de séries temporelles avec la même fréquence d'acquisition comme le clustering sur les séries 'hebdomadaires' de la filière. Dans le cas des données aquacoles, les valeurs de qualité de l'eau peuvent varier de manière significative, d'une semaine à l'autre, en cas de forte pluie par exemple. Ce niveau de résolution pour une analyse multi-variée, sur l'ensemble de la durée des élevages, n'est pas adapté; De plus, par rapport à ces variations hebdomadaires, Les phénomènes météorologiques présents dans la région comme par exem-

ple les dépressions tropicales et/ou de cyclones, peuvent modifier subitement la qualité de l'eau, sans que ce soit perceptible, dans la base de données, avec des relevés hebdomadaires des valeurs.

Les scénarios de l'analyse temporelles multi-variées et multi-échelles Comme pour l'analyse monovarié faite avec *X-meansTS*, le paramètre nb_min_clust (nombre de clusters initial) sera fixé à 5 (qui sera ensuite modifié selon la nouvelle mesure de dispersion) et le nombre minimum d'instances par cluster nb_min_inst sera fixé à 10. Pour rappel, nb_min_clust est fixé à 5 en raison des 5 typologies de croissance relevés dans le chapitre 6. Concernant le paramètre lié au seuil de dispersion (calculé par variable), rappelons que la recherche de ce seuil est automatisé pour chaque variable, à partir d'une liste de seuil obtenue sur des clusters générés par la méthode *K-meansDTW*. Notons qu'un seuil de dispersion pour une variable quelconque, aura un effet sur l'affinement général du clustering multi-varié car si pour une composante monovariée, d'un cluster multivarié, le seuil de dispersion (lié à la variable) n'est pas respecté alors, le cluster multivarié est re-partitionné.
On favorisera un seuil assez élevé, pour chaque variable, dans le cas de l'analyse multivariée, à la place du seuil moyen utilisé dans l'analyse mono-varié. Ainsi à partir de la liste de seuil obtenue pour une variable, le seuil maximal est conservé (cf section 7.2.4 pour les valeurs maximales par variable), pour l'analyse multi-variée présentée dans la section 7.3. En effet, dans la figure 7.13 on peut voir le résultat, obtenu avec un seuil moyen, d'un clustering multi-variés par la méthode *X-meansMMTS*, à partir de trois variables temporelles qui sont la température, l'oxygène dissous et le renouvellement d'eau (avec une fréquence uniquement journalière). Le seuil maximal est donc privilégié, pour les résultats de l'analyse multi-variée, car un seuil moyen, comme le montre la figure 7.13, peut générer des clusters avec très peu d'individus.

Clustering multi-varié multi-échelle sur les variables journalières et hebdomadaires avec un potentiel descriptif. Les variables considérées dans l'analyse multivariée et multi-échelles ci-dessous, sont toutes issues des séries temporelles de fréquence journalière et hebdomadaire. On y intégrera les séries de ces variables, issues de périodes pertinentes (i.e avec un potentiel descriptif de la performance d'élevage). D'après les résultats précédents de l'analyse mono-variée par *X-meansTS* (section 7.2), ces séries temporelles sont les suivantes :

- séries hebdomadaires : les 10 premières semaines de renouvellement d'eau et de salinité;
- les séries journalières : l'alimentation, la température et l'oxygène dissous avant le point d'inflexion de la courbe de croissance (on gardera les 50 premiers jours).

La figure D.1 montre le résultat sur ces séries de l'application de *XmeansMMTS*, avec un seuil élevé (pour chaque variable). De haut en bas, 5 clusters multivariés, regroupant les données d'élevages. Chaque cluster multivarié contient des élevages différents, regroupés par *XmeansMMTS* selon 9 variables temporelles de qualité de leur eau. Ces variables, affichés de gauche à droite, représentent l'évolution de :

- la température journalière,
- la température hebdomadaire,
- la salinité,
- le taux de renouvellement d'eau (hebdomadaire),
- la température avant le point d'inflexion de la croissance (50 premiers jours),
- le taux d'alimentation (journalier) avant le point d'inflexion de la croissance (50 premiers jours),
- le taux d'alimentation (journalier) sur l'ensemble de l'élevage (150 jours),
- le taux de renouvellement d'eau avant le point d'inflexion de la croissance (10 premières semaines),
- la salinité avant le point d'inflexion de la croissance (10 premières semaines),

Selon la figure 7.15, les distributions par cluster multiariés, avec un seuil moyen, sont plus homogènes (moins étendues) que celles de l'analyse mono-variée avec un seuil élevé. Ce constat est le même pour tous les clusters multi-variés, obtenus avec un seuil élevé (voir en Annexe D 'comparaison avec *K-Shape*'). L'analyse mono-variée présentée, dans la section précédente, a été réalisée à l'aide de seuils moyens. L'application de *X-meansMMTS* avec des seuils moyens entraîne également des distributions plus homogènes (sur les données de qualité production, par cluster), que celles de l'approche mono-variée avec le même seuil. Néanmoins, un seuil moyen par l'approche multi-variée génère des clusters avec trop peu d'individus.

Rappelons néanmoins, qu'en raison de l'imprécision des données (données manquantes), les résultats présentés visent à démontrer l'intérêt de l'approche multi-variée sur des données environnementales qui est de générer de nouveaux descripteurs (des représentants) fiables. Ces descripteurs sont générés par variable sur la base d'un nombre d'élevage limité. L'implication de ces descripeurs dans l'approche multivariée permet un clustering multivarié sur la base d'attributs représentatifs de l'évolution de chacune des variables.

Variable	Paramètre	cluster 0	cluster 3	cluster 9
Renouvellement d'eau les 10 premières semaines	coefficient de pente	1.578	1.447	2.349
	ordonnée à l'origine	8.114	5.579	8.187
Taux d'alimentation pris les 50 premiers jours	coefficient de pente	0.107	0.095	0.128
	ordonnée à l'origine	0.416	0.505	0.509

Table 7.5 Paramètres des modèles de régression, des représentants des taux de renouvellement d'eau, pris sur les 10 premières semaines d'élevage et de l'alimentation sur les 50 premiers jours.

Les distributions affichées ont par pairs de clusters des p-valeurs inférieurs à 0.05. L'approche multi-variée a extrait, contrairement à l'approche mono-variée, des pairs de clusters avec des distributions significatives de taux de branchies oranges.

Par exemple, malgré le faible nombre d'individus pour le cluster 9, l'approche a regroupé les élevages contenant un taux de défauts sur les branchies relativement plus élevé que les autres clusters. Malgré ce faible nombre d'individus, la figure 7.15, confirme que ce défaut n'impacte pas la survie des espèces, puisque le cluster 9 (contenant le taux de branchies oranges le plus élevé), a aussi le taux de survie le plus élevé [106, 105]. Il impacterait potentiellement, la croissance initiale de l'animal (paramètre C de Gompertz), une augmentation de cette vitesse (avant l'inflexion de la courbe de croissance) correspond avec un taux de branchies oranges relativement plus important. En effet les pairs de clusters dont les distributions de défauts de branchies, ont eu une p-valeur inférieure à 0.05 avaient également, des distributions de vitesse de croissance initiale significativement différentes.

Les informations sur les représentants des clusters sont affichées (figure 7.16). Le tableau 7.5 les paramètres des modèles de régression, des représentants des taux de renouvellement d'eau, pris sur les 10 premières semaines d'élevage et de l'alimentation sur les 50 premiers jours.

Cette figure et ce tableau, montrent que lorsque les taux de renouvellement et d'alimentation sont importants, cela favorise une importante accélération de la vitesse initiale. Cette accélération est plus importante que l'accélération induite par une température plus élevée en début d'élevage, comme c'est le cas pour le cluster 3 si on le compare au cluster 9.

7.3.0.1 Classification non supervisée des bassins, par la méthode *X-meansMTS* en fonction de la qualité du milieu

En considérant l'ensemble des élevages analysées par *XmeansTS*, l'approche multivariée, a permit de regrouper des productions, ensemencés à des périodes identiques, et qui sont liés à un seul identifiant de ferme. La figure 7.17 montre pour les bassins d'une même ferme, que la méthode a regroupé des élevages ensemencés entre les mois

d'août et septembre (cluster 0) et des élevages ensemencés entre octobre et décembre (cluster 8). Ces élevages ont été réalisés entre 2010. Le mois d'ensemencement était la variable explicative prépondérante dans l'analyse mono-variée. L'approche multivariée permet d'affiner le modèle descriptif, d'extraire des caractéristiques plus localisées, des tendances de l'évolution de la qualité du milieu relatives aux productions d'une même ferme. L'approche mono-variée, extrait des tendances qui sont présentes sur plusieurs élevages sans distinctions de fermes, et donc liées potentiellement à des normes de performances à l'échelle de la filière.

7.4 Synthèse de l'application de l'approche multi-variée sur les données aquacoles

L'enjeux de l'analyse multi-variée a été d'améliorer la compréhension des corrélations entre l'impact conjugué de différentes variables temporelles, sur les distributions des données de performance. L'approche multi-variée a permis d'identifier des groupes de clusters multivariés, à certains élevages réalisés au sein d'une même ferme. Il s'agit ensuite de déterminer si l'état du bassin (données édaphiques...) impacte considérablement la qualité finale de la production. Cette méthode est en soit une contribution majeure dans le domaine des sciences de données, et permettra, d'être appliquée à un ensemble de données d'autres filières de production aquacoles et agricoles.

Enfin notons que la distribution des données de performance par cluster, a montré que les paramètres de productivité (croissance et survie) sont davantage corrélés à la qualité du milieu durant la période de grossissement, que les défauts qui apparaissent durant cette même période (et qui sont relevés par la SOPAC en fin de production). Ce constat n'induit pas une non-corrélation entre les défauts et la qualité du milieu. Il met en avant que l'étude des défauts sur une période de grossissement ne permet pas d'expliquer l'apparition de ces défauts. Et que par exemple une analyse descriptifs de l'évolution des paramètres physico-chimique, édaphique d'un même bassin doit être faite sur plusieurs élevages consécutifs.

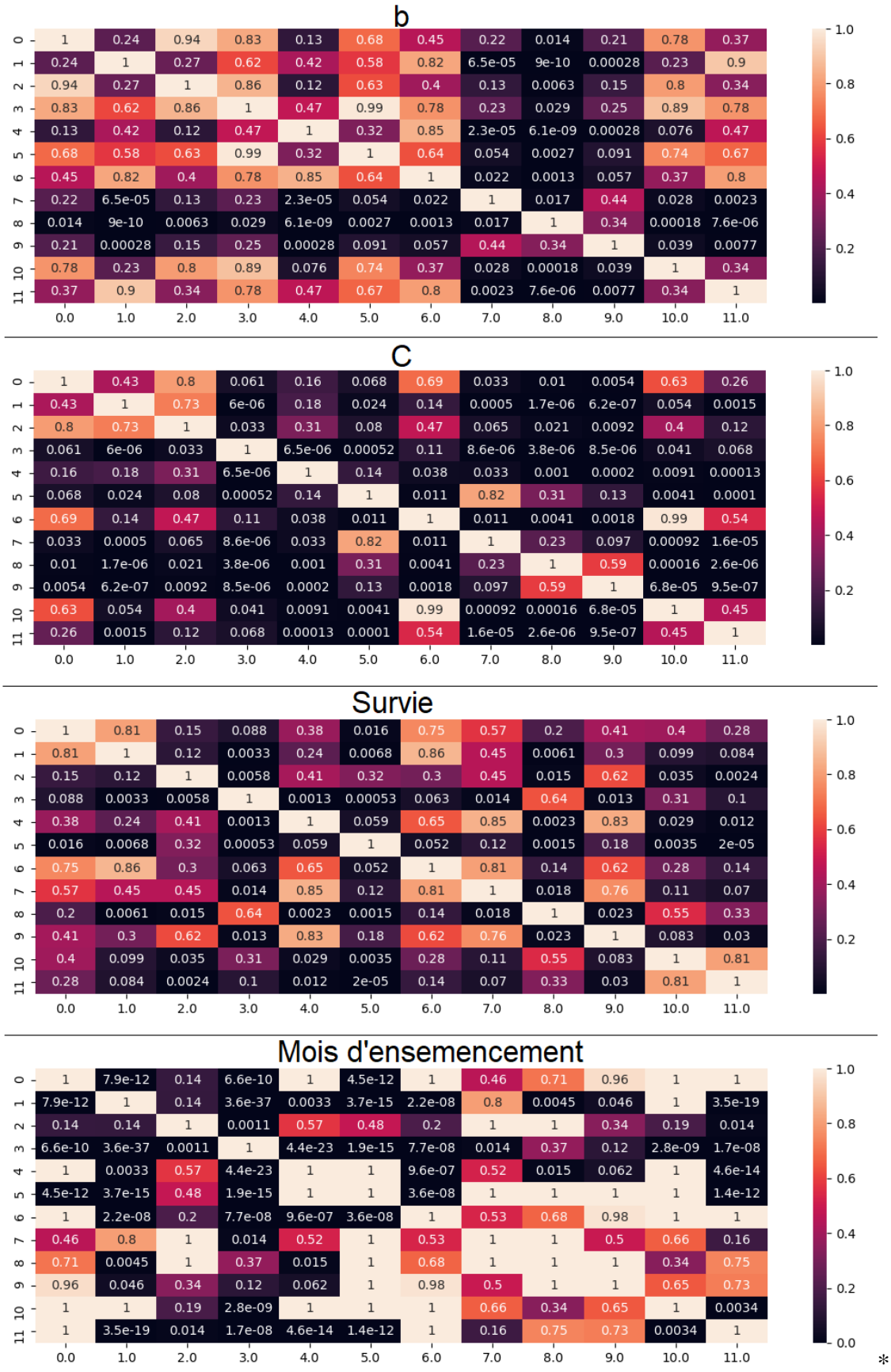

Fig. 7.6 Distribution de la croissance initiale *C*, la vitesse de convergence vers le poids final *B*, la survie et le mois d'ensemencement, par pairs de clusters de température et obtenus par la méthode XmeanTS

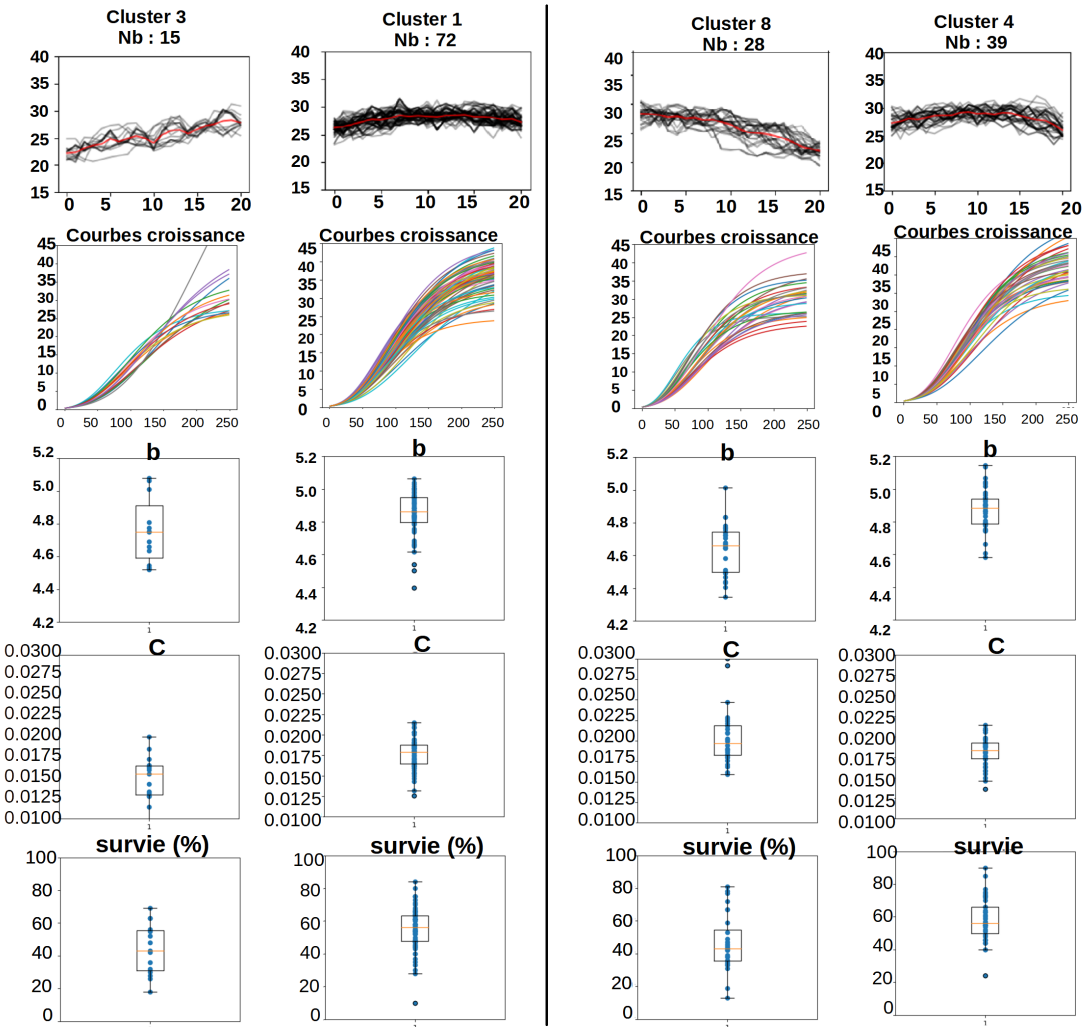

Fig. 7.7 Distribution de la croissance initiale (C), de la vitesse de convergence et de la survie en fonction des clusters des températures hebdomadaires, avec des *p-valeurs* inférieurs à 0.05% avec la méthode *Xmeans-TS*.

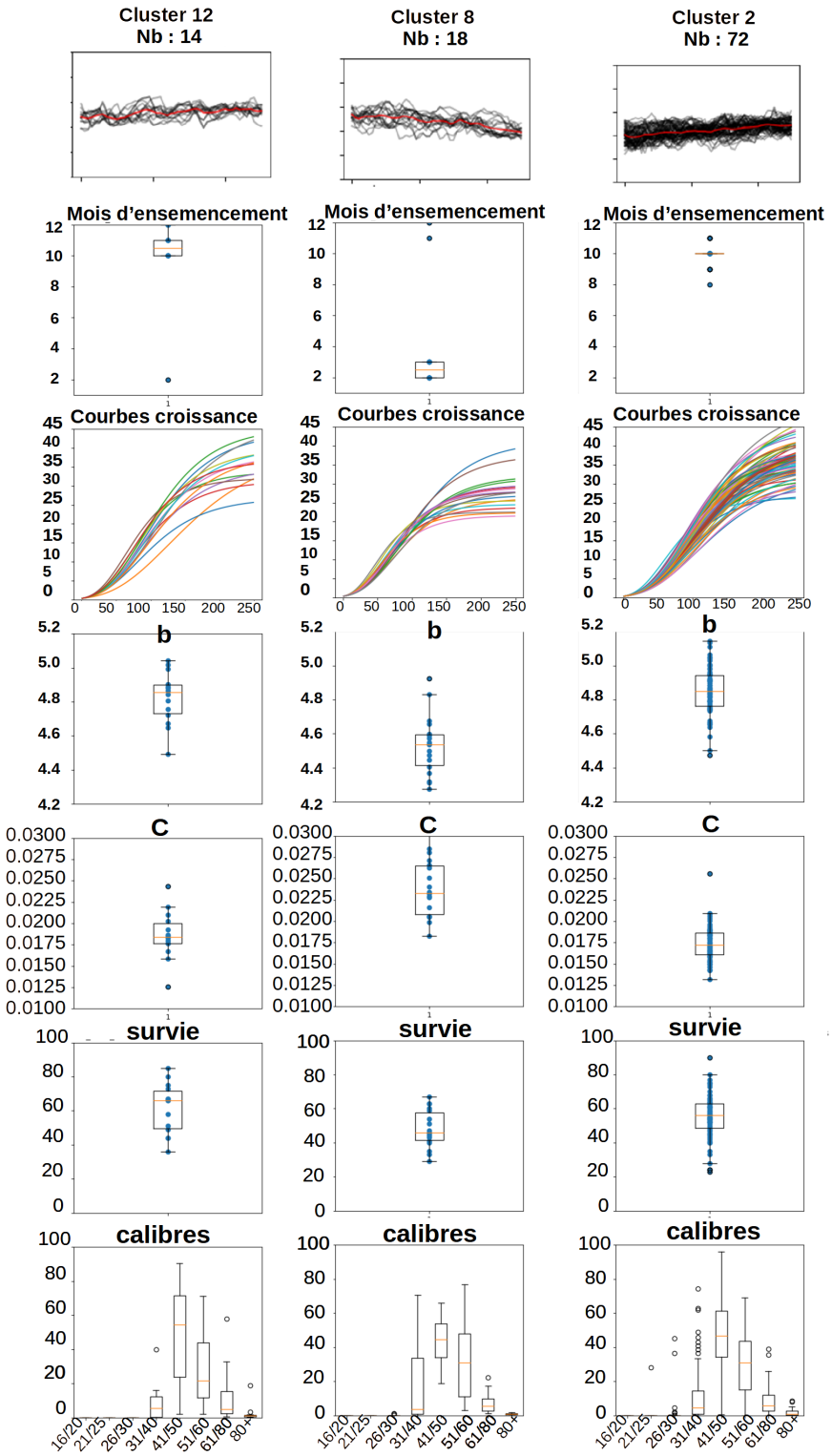

Fig. 7.8 Distribution de la survie en fonction des clusters de température journalière ayant les p-valeurs inférieur à 0.05 pour *Xmeans-TS*

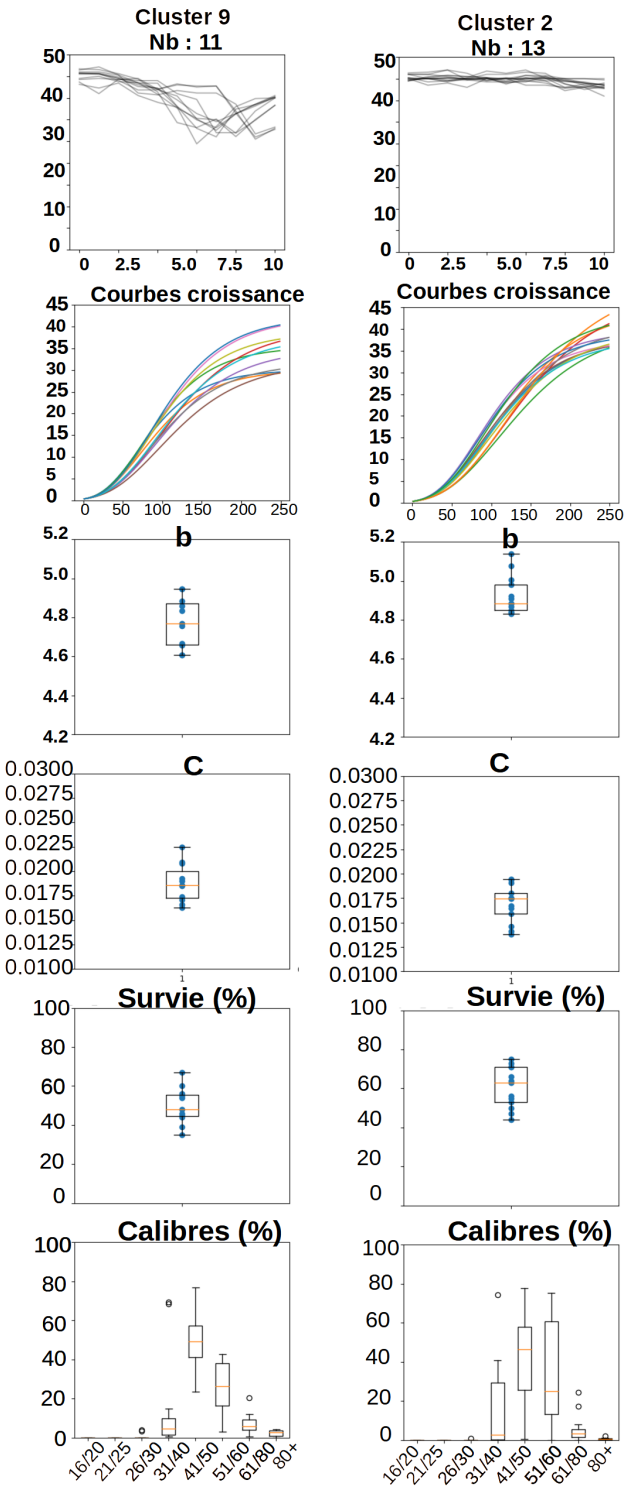

Fig. 7.9 Clusters de la salinité obtenus par la méthode *X-meanTS*

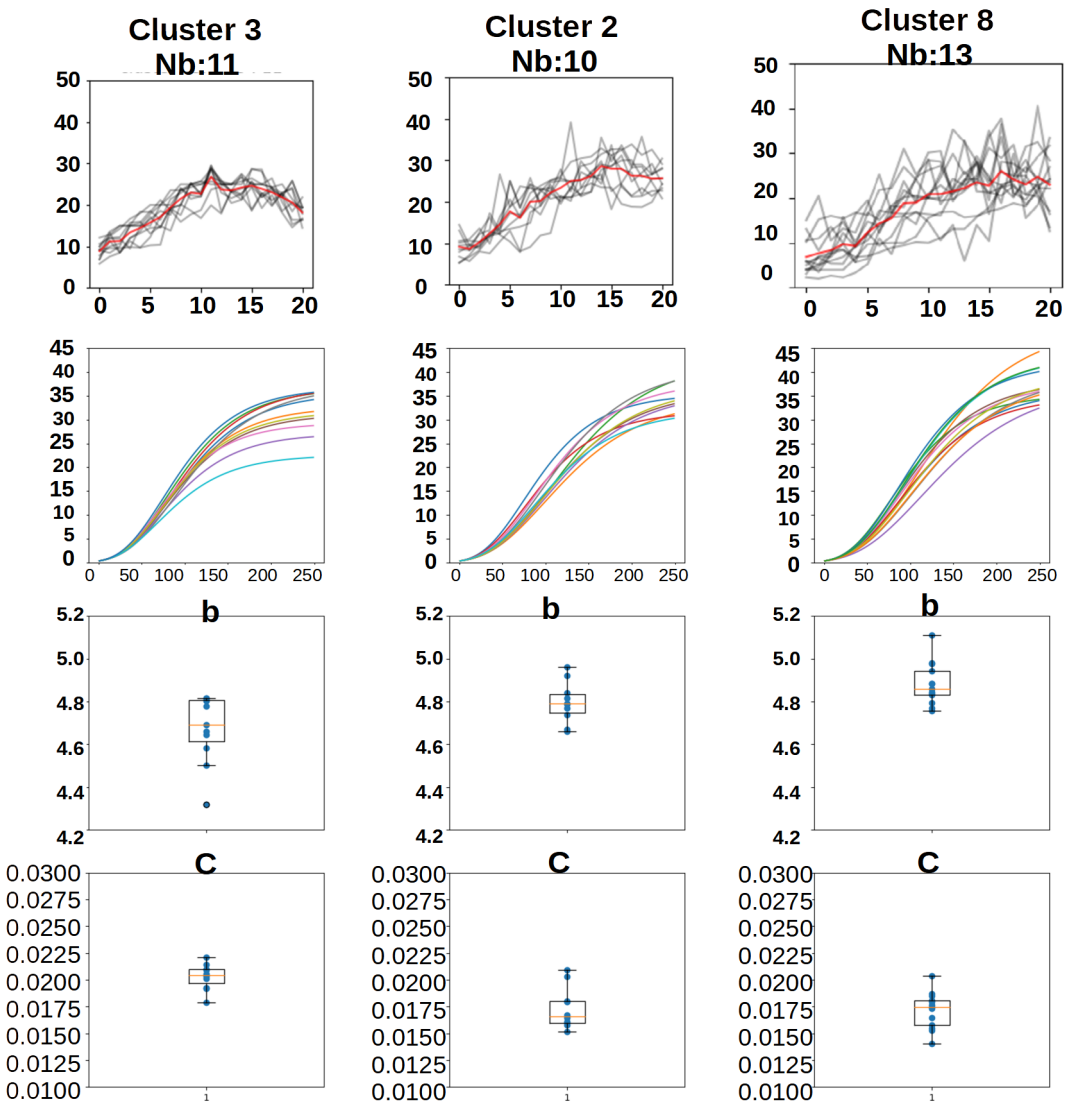

Fig. 7.10 Distribution de la vitesse, initiale et de la vitesse de convergence en fonction des clusters 3, 2 et 8 de renouvellement d'eau, obtenus avec un seuil de distribution minimal

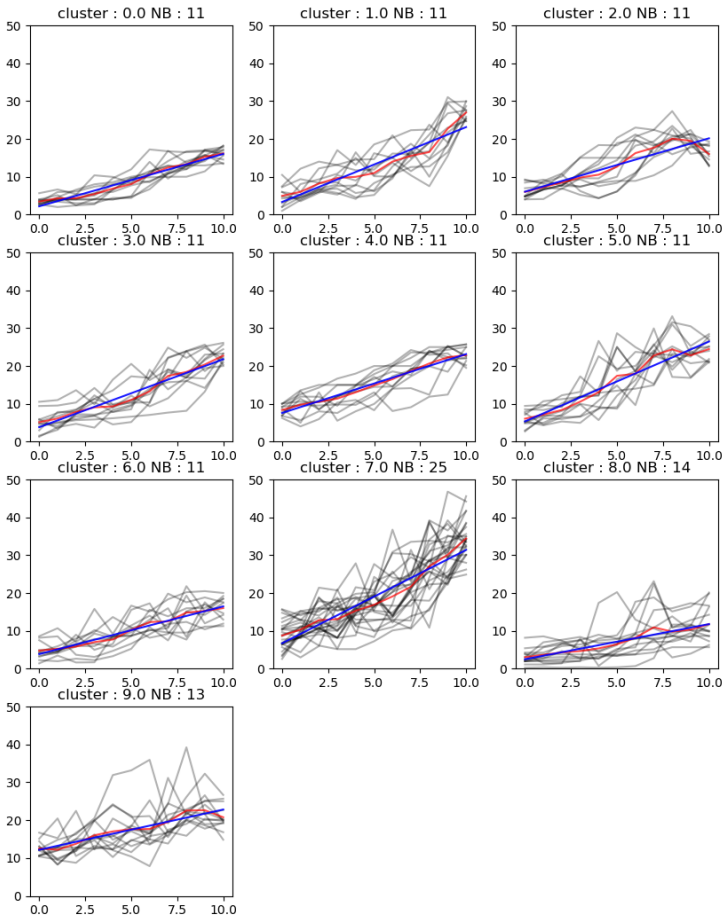

Fig. 7.11 Cluster de renouvellement de l'eau prise sur les 10 premières semaines d'élevage

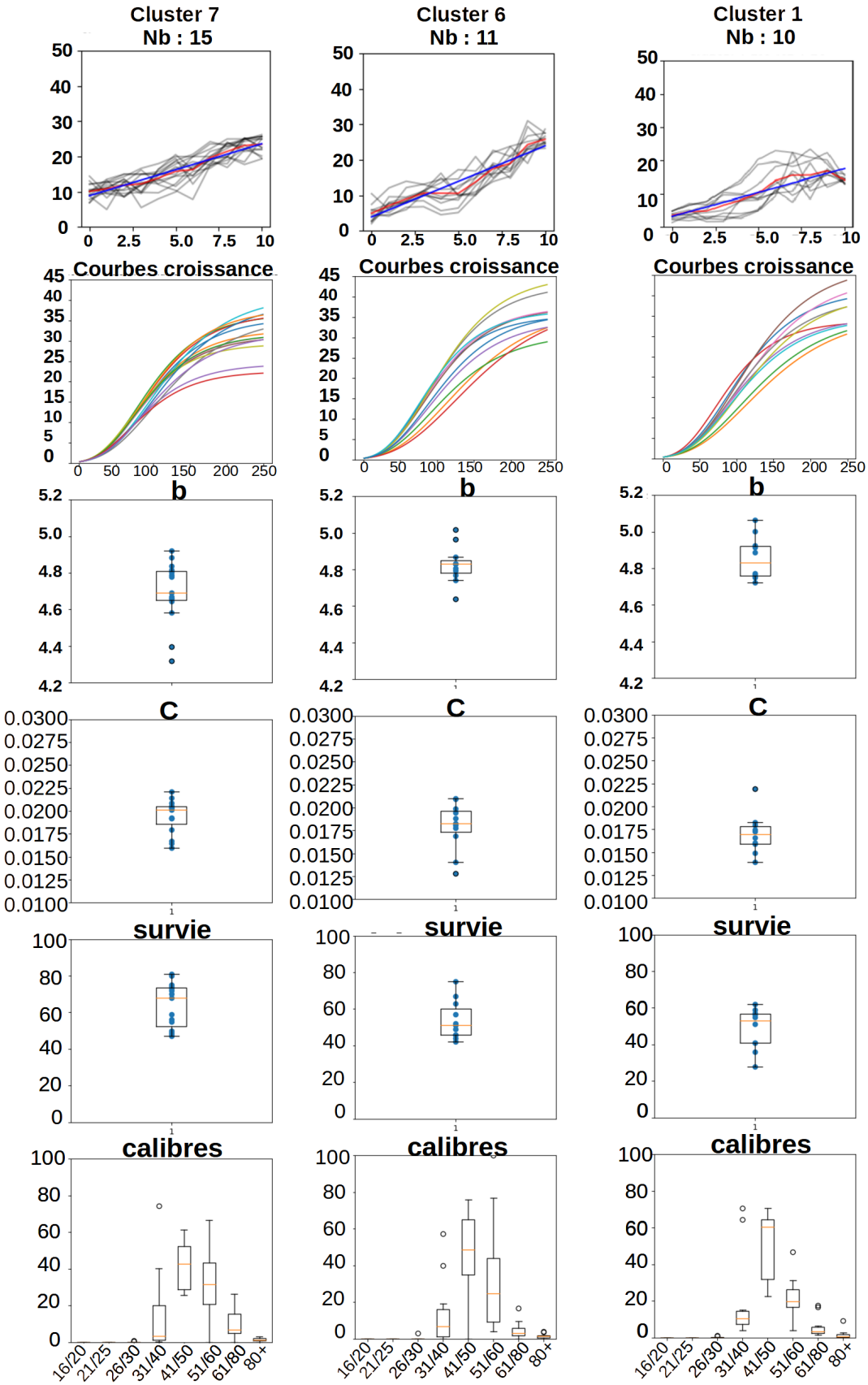

Fig. 7.12 Distribution de la vitesse de convergence de convergence vers le poids final, de la vitesse de converge et de la survie en fonction des clusters 7, 6 et 1 de renouvellement d'eau durant les 10 premières semaines d'élevage, obtenus avec un seuil de distribution minimal

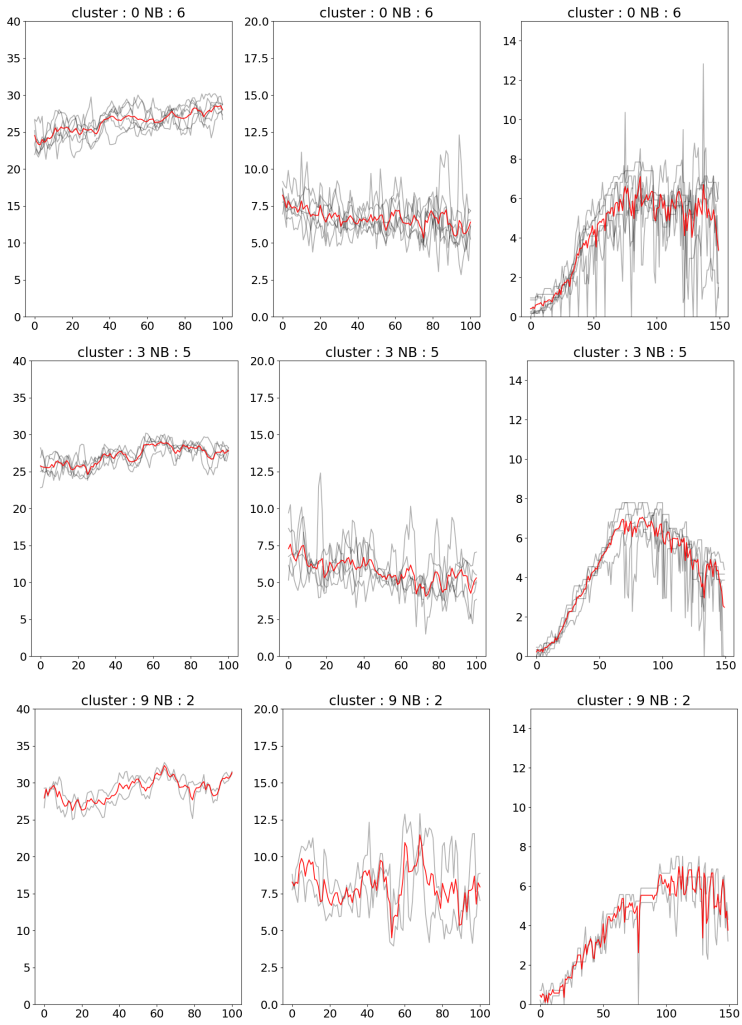

Fig. 7.13 Clustering multivarié de variables avec une fréquence journalière et un seuil de dispersion faible

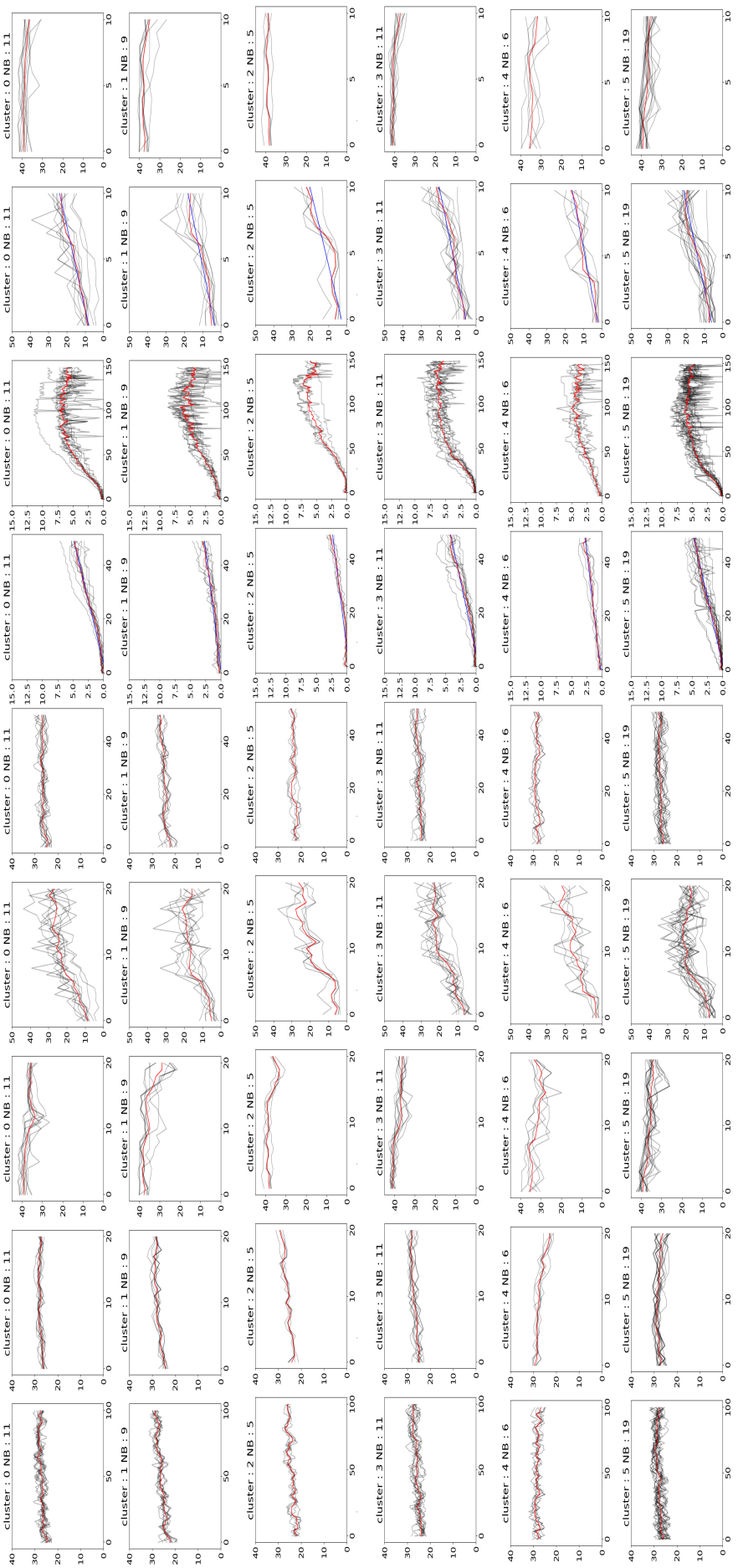

Fig. 7.14 Clustering multivariée multi-échelle

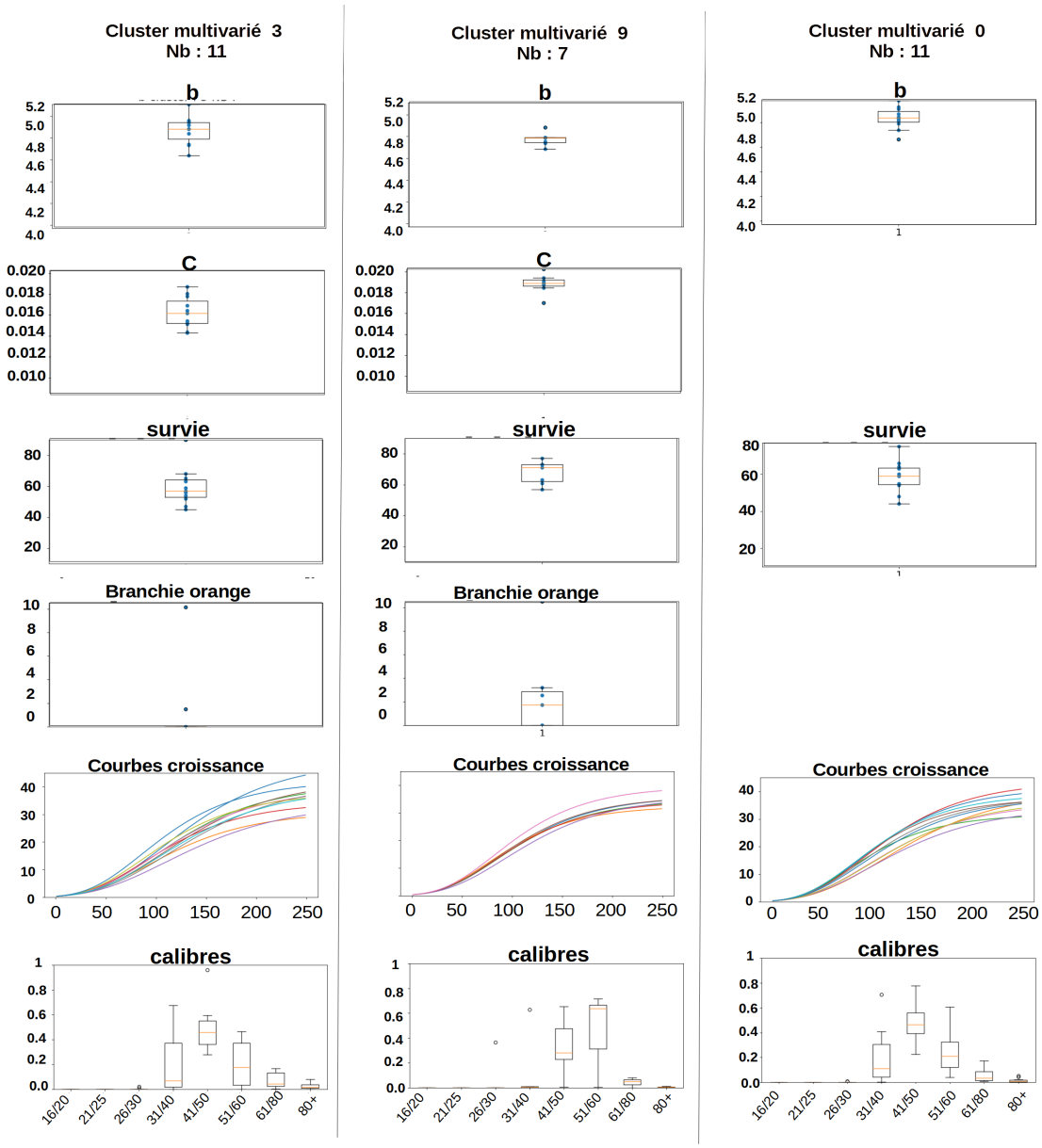

Fig. 7.15 Clustering multivariée multi-échelle

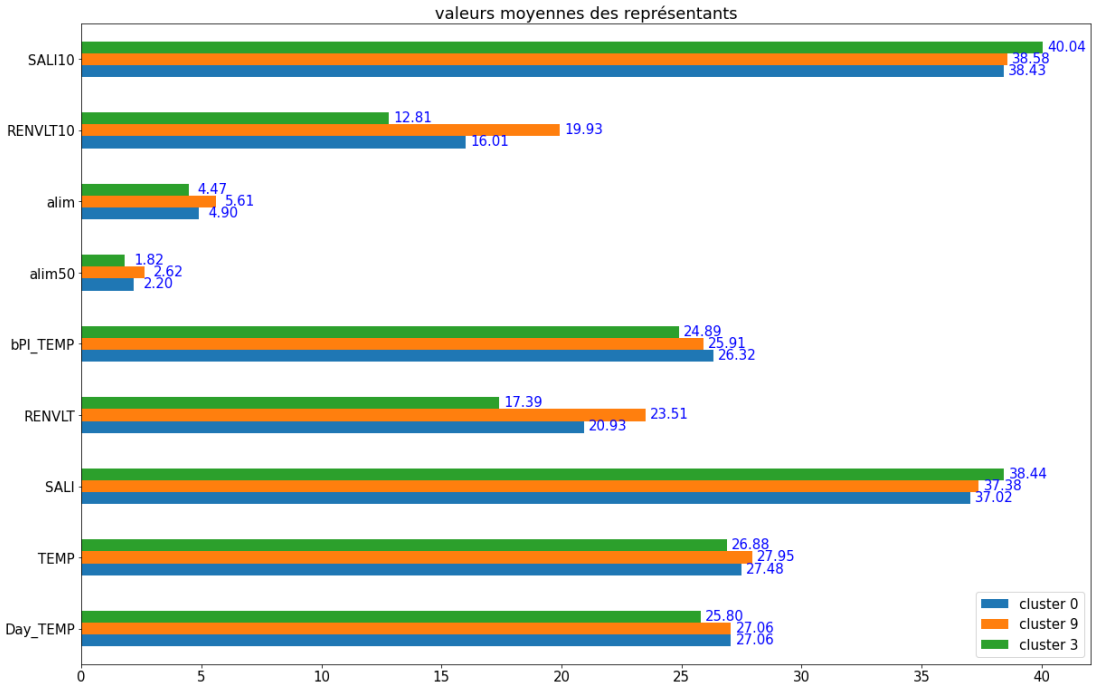

Fig. 7.16 Valeur moyenne des représentants par variable, pour les clusters 0, 9 et 3 générés par un *X-meanMMTS*

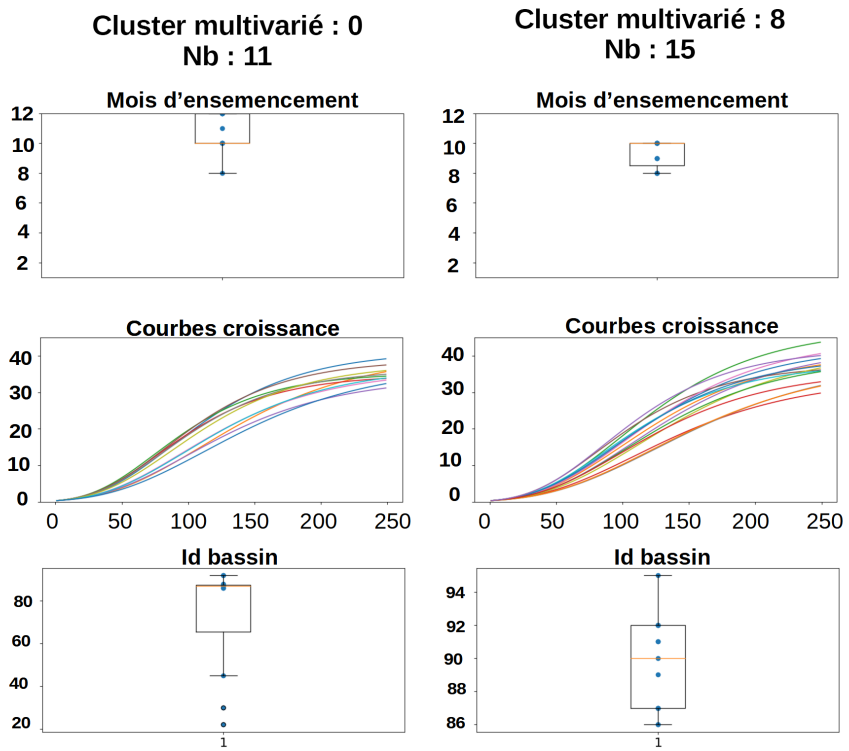

Fig. 7.17 classification non supervisée des bassins d'une même ferme, par la méthode *X-meansMMTS* en fonction de la qualité du milieu

Conclusion

Nous nous sommes intéressé dans cette thèse à l'utilisation des méthodes en science de données pour extraire de nouvelles connaissance à partir des données générées dans les systèmes de production aquacoles. Dans le domaine aquacole et agricole, l'analyse en amont des paramètres zootechniques, permet d'extraire des informations pertinentes à l'échelle d'une filière, et notamment des typologies de croissance. L'analyse des données de qualité du milieu d'élevage, en fonction de l'évolution de formes et d'amplitudes des variables temporelles concernant l'évolution de cette qualité, a été possible à partir des nouvelles approches proposées dans le chapitre 4. Ces approches sont des méthodes de clustering de séries temporelles mono-variées *Xmeans-TS* [167], et multi-variées, multi-échelles *X-meansMMTS* [168]). La stratégie d'analyse de données de la filière aquacole étudiée, et détaillée dans le chapitre 5, (présentée durant un workshop sur la Gestion et l'Analyse des données Spatiales et Temporelles (GAST) à Paris), a été relevée comme étant structurée et pertinente, par les experts en sciences de données. La stratégie d'analyse de l'ensemble des données générées dans la filière a permit d'associer les typologies de croissances aux évolutions de qualité de l'eau des productions. L'analyse mono-variée a généré des clusters avec des caractéristiques globales i.e des tendances générales, présentes dans les séries de données de qualité d'eau à l'échelle de la filière. L'analyse multi-variée a permit d'affiner ces tendances, et a extrait des clusters relatifs aux élevages appartenant à une ferme; Elle permet de regrouper des productions en fonction des bassins de la ferme. Elle a mit en évidence d'autres liens, par exemple entre le défaut de branchies et la vitesse de croissance initiale, et qui n'étaient pas identifiables par l'analyse mono-variée. Les deux approches ont leurs intérêts, tant pour le domaine que pour la science de données. L'approche multi-variée est décrite comme une contribution majeure en soit, et les possibilités d'application dans différents domaines sont variés.

7.5 Amélioration de la méthodologie d'extraction de connaissance dans les filières aquacoles

L'intégration des méthodes crées (*X-meansTS* et *X-meansMMTS*) dans un processus complet d'extraction de connaissance dans des données (*ECD* Cf chapitre 1 section 1.2.1), aidera les aquaculteurs, à avoir une compréhension plus fine des relations entre la qualité du milieu, et la productivité. Néanmoins un travail important d'acquisition de données reste à faire. Des données importantes, non présentes dans la base de données *Stylibase*, n'ont pu être intégrées. Elles concernent des informations sur le sol, i.e des paramètres édaphiques, qui nous auraient permis d'analyser l'impacte de ces paramètres (édaphiques) sur la qualité de l'eau, après remplissage du bassin. Une autre donnée importante et sur laquelle, l'éleveur ne peut agir durant la phase de grossissement, est la génétique. Celle ci permettra de comprendre si l'influence des différents paramètres

zootechniques (vitesse de croissance initiale....), par la qualité du milieu, est fonction des caractéristiques génétiques des espèces.

Perspectives concernant l'acquisition de données : Pour rappel, le processus d'extraction de connaissances débute avec l'étape d'acquisition de données. Durant les différentes phases de grossissement de l'espèce, l'acquisition de données par des capteurs ou par des images, aiderait les fermiers et les sociétés de transformations telles que la *SOPAC*, à préciser les taux des différents défauts présents sur une production, ou encore les aspects caractéristiques de la présence d'une maladie. En effet, comme énoncé, dans le second chapitre, des techniques de reconnaissances de maladies (reconnaissance de caractéristiques visuelles) ont été appliquées dans le domaine agricole, sur les plantes, par exemple... Dans le domaine aquacole, l'acquisition de données est plus complexe, en raison de la colonne d'eau, qui est un obstacle à la transmission d'informations, et apporte du biais, en raison par exemple de la turbidité suite à l'accumulation de matières organiques. Néanmoins les techniques de correction (de ce biais) ont montré que la reconnaissance automatique (des espèces) atteignait d'excellentes performances, pour l'extraction de caractéristiques zootechniques des espèces étudiées. Ainsi, la reconnaissance de la présence de défauts de production (branchies oranges, cicatrices ...) dans la filière crevetticole calédonienne, peut être détectées dès son apparition, durant la phase de grossissement, par l'analyse d'images prises sous l'eau. L'analyse d'images de crevette par des réseaux connexionistes, doit être réalisée, afin d'apprendre d'une part à classifier les types de défauts. Mais elle doit permettre aussi de reconnaître de nouveaux défauts. Cela permettra d'avoir enfin des outils visant à anticiper l'expansion de certaines maladies sur l'ensemble des productions d'une ferme, ou de la filière.

Au niveau de l'acquisition, les protocoles de suivis de la qualité du milieu, par les aquaculteurs, doivent être normalisés sur l'ensemble des fermes de la filière. Les fréquences d'enregistrement des données, doivent être fonction, de l'expertises terrains, et des résultats obtenues sur l'analyse mono-variées et muti-variées des variables temporelles de qualité du milieu. Par exemple, l'approche *X-meansTS* a mis en évidence des niveaux de résolutions pertinents, en fonction de différentes variables de qualité d'eau des bassins (température, salinité...).
L'acquisition de données de qualité d'eau à différentes fréquences permet aux modèles d'analyses multi-échelles, d'identifier les périodes pertinentes, et les différents niveaux de résolution, avec un fort potentiel descriptif et prédictif de la qualité de production.

7.5.0.0.1 Perspectives liées au traitement des données : L'évolution de la qualité du milieu de production de ressources naturelles, génère des séries temporelles de données complexes, i.e avec des variations importantes. Une étude comparative des méthodes permettant de réduire cette complexité, en amont de l'utilisation des méthodes crées (*X-meansTS* et *X-meansMMTS*), serait à faire. Par exemple les méthodes de lissage, appliquées aux séries d'oxygènes dissous qui ont de fortes variations dans le temps, permettraient d'extraire d'avantage de clusters, avec des tendances homogènes. Cependant ces méthodes pourraient conduire à une perte d'informations. Cette perte doit être prise en compte, car les variations influent considérablement sur les espèces, et donc sur la productivité des filières.

Les paramètres en entrée des méthodes proposées (*X-meansTS* et *X-meansMTS*), déterminent la dispersion des distances entre les séries temporelles analysées et leur représentant. Pour rappel, la méthode *X-meansTS* utilise des méthodes existantes de clustering de séries temporelles qui se servent d'une distance adaptée à la comparaison de deux séries. *X-meansTS* applique à cette distance, la mesure de dispersion. Cette mesure peut être appliquée à toutes les méthodes de clustering basées sur des mesures de distances. La recherche automatique du seuil critique pour cette mesure peut être améliorée, en vu d'optimiser l'homogénéité finale (de l'ensemble des clusters). Par exemple, l'approche de clustering *DBScan*, identifie les individus, à partir desquels, la recherche de groupe débute. La densité d'individus, qui seront inclus dans le même cluster, est relativement plus important, autour de ces individus sources. L'amélioration à apporter à *X-meansTS*, peut être d'identifier, sur l'ensemble du jeu, les groupes de séries, et qui, par exemple, ont une mesure de dispersion, plus élevée, par rapport à leur série moyenne.

La mesure de dispersion peut être directement adaptée à la méthode existante *K-means*, de clustering de données statiques. L'intégration de cette mesure, dans cette méthode, est une approche en cour de réalisation. Cette approche dérivée de *K-means* peut donc être inscrite dans la nouvelle méthode multi-variée *X-meansMTS*. En effet la méthode multi-variée génère une matrice de données statiques, sur laquelle est appliquée *K-means* de manière hiérarchique.

L'enjeux en science de données, énoncé dans le chapitre 5, est aussi de superviser les modèles descriptifs mono-variées et multi-variées proposées. Pour cela, nous pouvons nous intéresser aux approches d'apprentissages supervisées basées sur une exploitation de distributions Gaussiennes des données. En effet la mesure de dispersion est adaptée, à ce type de distribution. L'enjeux est de créer une méthode de classification supervisée de séries temporelles multi-variées et multi-échelles qui soit inter-prétable. En effet, il existe actuellement aucune méthode qui réponde à cette problématique et qui restitue un modèle d'apprentissage interprétable;

Une analyse (i.e inexistante dans la littérature), concernant le calcul de l'homogénéité des clusters multi-variés doit être réalisée. Cette analyse est possible en utilisant des mesures de performances dédiées aux clustering. Pour cela, *X-meansMMTS* peut être appliqué à des combinaisons de variables de qualité du milieu d'élevage. Les clusters multivariés peuvent être décrits par des cibles générées à partir de combinaisons de variables de performance d'élevage. En effet, la mesure de dispersion intégrée dans l'algorithme *MMTS*, permet d'augmenter l'homogénéité des clusters. L'objectif est de comparer l'homogénéité des clustering générés sur ces combinaisons de séries.

7.5.0.0.2 Une analyse des données temporelles avec des niveaux de résolutions hétérogènes : D'autres variables temporelle peuvent être créées à partir des séries exploitées, afin d'analyser leurs évolutions à différentes échelles temporelles; ci-dessous, par exemple, sont décrits certaines de ces variables à créer :

- Température et oxygène dissous journaliers, du matin :
  (en faisant par ex la moyenne des valeurs prises entre 6h et 8h)
- Température et oxygène dissous journaliers du soir :
  (en faisant par ex la moyenne des valeurs prises entre 16h et 20h)
- Différence de quantité d'aliment distribué entre deux jours consécutifs
- ....

A partir de ces variables, des scénarios de combinaison d'analyse multi-variée, c'est à dire comprenant 2 à N variables temporelles, avec N le nombre de variable de qualité du milieu, sont à réaliser et à interpréter. L'objectif étant d'évaluer l'homogénéité des clusters, calculé sur les distributions de données de productivités, en modifiant les variables temporelles utilisées et leurs niveaux de résolutions. Des premiers tests effectués ont montrés que l'analyse conjuguée de la température, et de l'oxygène dissous du soir avant l'inflexion de la croissance, créaient des clusters plus homogènes, par rapport à la vitesse de croissance initiale et la survie.

Une méthodologie de hiérarchisation de l'influence des variables temporelles utilisées sur des cibles, est une perspective à l'approche multivariée proposée. Pour cela l'approche doit donc être supervisée. Et de manière équivoque au modèle supervisé d'arbre de décision, qui hiérarchise les attributs en fonction de leurs niveaux dans l'arbre, cette restitution permettrait d'interpréter plus facilement l'analyse multivariée et multi-échelle. La hiérarchisation des attributs du nouveau modèle à superviser, pourrait être basé sur la mesure de dispersion des données, par variable.

L'extraction de connaissance à différentes échelles de temps Les réseaux de neurones comme le réseau *LSTM* [84], sont adaptées aux données temporelles multi-variées. Sans avoir obligatoirement à superviser l'approche *X-meansMMTS*, il est possible de tester des modèles supervisés sur la matrice de données, générées par cette approche. Ainsi, en testant différents classifiers avec cette matrice (comme jeu de donnée d'apprentissage), il serait envisageable de valider l'intérêt d'apprendre à partir des nouveaux descripteurs (représentant de l'ensemble des variables) crées par *X-meansMMTS*. Ceci est possible par comparaison de la précision obtenus par les classifiers avec ceux, par exemple, d'un réseau de neurones adaptés aux séries chronologiques, comme le réseau *LSTM*, testé sur les données brutes (les séries temporelles multi-variées). Cela validerait donc la pertinence de ces descripteurs et leurs potentiels prédictifs. Une première analyse effectuée sur un jeu de données en ligne, a montré que la performance s'améliorait en utilisant les données de la matrice, en comparaison à l'utilisation de *LSTM*, sur le jeu brute, dont la précision était de 93%. En utilisant la matrice, la performance de l'arbre de décision, avec la matrice, et du modèle *LDA* était de plus de 95%.

La restitution des nouvelles connaissances Enfin la phase de restitution des modèles doit être faite dans des interfaces ergonomiques, pour l'utilisateur, qui n'est pas familier avec les méthodes en science de données. Les clusters sont parfois difficilement interprétables (on le voit avec les clusters d'oxygène dissous). De plus, des caractéristiques communes dans les séries, intra-cluster, doivent être affichées (de manière ergonomique) et analysées. Cela permettra aux aquaculteurs, d'expliquer les distributions de données de productivités, par des évolutions particulières de paramètres de qualités du milieu i.e des sous-ensembles de données de qualité du milieu contenant des caractéristiques spécifiques, pertinentes.

Références

[1] A Knowledge-Based Approach for Supporting Aquaculture Data Analysis Proficiency, volume Volume 2B: Advanced Manufacturing of ASME International Mechanical Engineering Congress and Exposition, 11 2015. doi: 10.1115/IMECE2015-52183. URL https://doi.org/10.1115/IMECE2015-52183. V02BT02A025.

[2] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. Information Systems, 53:16–38, 2015. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2015.04.007. URL https://www.sciencedirect.com/science/article/pii/S0306437915000733.

[3] Naji Al-Dosary, Saad Alhamed, and Abdulwahed Aboukarima. K-nearest neighbors method for prediction of fuel consumption in tractor-chisel plow systems. Engenharia Agrícola, 39:729–736, 12 2019. doi: 10.1590/1809-4430-eng.agric.v39n6p729-736/2019.

[4] Suraj Amatya, Manoj Karkee, Aleana Gongal, Qin Zhang, and Matthew D Whiting. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. Biosystems engineering, 146:3–15, 2016.

[5] Duong Tuan Anh and Le Huu Thanh. An efficient implementation of k-means clustering for time series data with dtw distance. International Journal of Business Intelligence and Data Mining, 10(3):213–232, 2015.

[6] BENYETTOU Assia. Contribution en apprentissage semi-supervise sous contexte multi-label. PhD thesis, Oran, 2017-2018.

[7] Anthony J. Bagnall, pages=43-49 Hoang Anh Dau volume=26, number=1, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn J. Keogh. The UEA multivariate time series classification archive, 2018. CoRR, abs/1811.00075, 2018. URL http://arxiv.org/abs/1811.00075.

[8] Ziv Bar-Joseph, Georg Gerber, David Gifford, Tommi Jaakkola, and Itamar Simon. A new approach to analyzing gene expression time series data. Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB, 02 2002. doi: 10.1145/565196.565202.

[9] Lefteris Benos, Aristotelis Tagarakis, Georgios Dolias, Berruto Remigio, Dimitrios Kateris, and Dionysis Bochtis. Machine learning in agriculture: A comprehensive updated review. Sensors, 21, 05 2021. doi: 10.3390/s21113758.

[10] Donald J. Berndt and James Clifford. Finding patterns in time series: A dynamic programming approach. In Advances in Knowledge Discovery and Data Mining, pages 229–248. 1996.

[11] Heinz Bernhardt, Sebastina Götz, Nina Zimmermann, and Dirk Engelhardt. Simulation of agricultural logistic processes with k-nearest neighbors algorithm. Agricultural Engineering International : The CIGR e-journal, 2015:241–245, 05 2015.

[12] J Martin Bland and Douglas G Altman. Statistics notes: Measurement error and correlation coefficients. BMJ, 313(7048):41–42, jul 1996.

[13] Gerry Bourke, Frank Stagnitti, and Brad Mitchell. A decision support system for aquaculture research and management. Aquacultural Engineering, 12(2):111 – 123, 1993. ISSN 0144-8609. doi: https://doi.org/10.1016/0144-8609(93)90020-C. URL http://www.sciencedirect.com/science/article/pii/014486099390020C.

[14] Claude Boyd and Sergio Zimmermann. Grow-out systems-water quality and soil management. Freshwater prawns: Biology and farming, pages 239–255, 2010.

[15] Leo Breiman. Bagging predictors. Machine learning, 24(2):123–140, 1996.

[16] Glaucia M. Bressan, Vilma A. Oliveira, Estevan R. Hruschka, and Maria C. Nicoletti. Biomass based weed-crop competitiveness classification using bayesian networks. In Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), pages 121–126, 2007. doi: 10.1109/ISDA.2007.60.

[17] David Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. ROYAL SIGNALS AND RADAR ESTABLISHMENT MALVERN (UNITED KINGDOM), RSRE-MEMO-4148, 03 1988.

[18] M Burford. Phytoplankton dynamics in shrimp ponds. Aquaculture Research, 28(5):351–360, 1997.

[19] Debouche Charles. Presentation coordonnée de differents modeles de croissance. Revue de Statistique Appliquee, pages 5–22, 1979.

[20] Lei Chen and Raymond T. Ng. On the marriage of lp-norms and edit distance. In VLDB, pages 792–803, 2004. ISBN 0-12-088469-0.

[21] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

[22] YingYi Chen, YanJun Cheng, QianQian Cheng, HuiHui Yu, DaoLiang Li, et al. Short-term prediction model for ammonia nitrogen in aquaculture pond water based on optimized lssvm. International Agricultural Engineering Journal, 26 (3):416–427, 2017.

[23] Zhi Cheng, Frédéric Flouvat, and Nazha Selmaoui-Folcher. Mining recurrent patterns in a dynamic attributed graph. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 631–643. Springer, 2017.

[24] Harshit Singh Chhabra, Akshay Kumar Srivastava, and Rahul Nijhawan. A hybrid deep learning approach for automatic fish classification. In Pradeep Kumar Singh, Bijaya Ketan Panigrahi, Nagender Kumar Suryadevara, Sudhir Kumar Sharma, and Amit Prakash Singh, editors, Proceedings of ICETIT 2019, pages 427–436, Cham, 2020. Springer International Publishing. ISBN 978-3-030-30577-2.

[25] ByoungSeon Choi. ARMA model identification. Springer Science and Business Media, 2012.

[26] Jesper Haahr Christensen, Lars Valdemar Mogensen, Roberto Galeazzi, and Jens Christian Andersen. Detection, localization and classification of fish and fish species in poor conditions using convolutional neural networks. In 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV), pages 1–6. IEEE, 2018.

[27] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. Mathematics of Computation, 19:297–301, 1965.

[28] D. Coomans and D.L. Massart. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules. Analytica Chimica Acta,

136:15–27, 1982. ISSN 0003-2670. doi: https://doi.org/10.1016/S0003-2670(01)95359-0. URL https://www.sciencedirect.com/science/article/pii/S0003267001953590.

[29] Randall L. Dahling. Shannon's Information Theory: The Spread of an Idea. In Wilbur Schramm, editor, Studies of Innovation and of Communication to the Public, pages 118–139. Stanford University Press, Stanford, 1962.

[30] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74:902–924, 2017.

[31] B Vikram Deep and Ratnakar Dash. Underwater fish species recognition using deep learning techniques. In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), pages 665–669, 2019. doi: 10.1109/SPIN.2019.8711657.

[32] Krzysztof Dembczyński, Weiwei Cheng, and Eyke Hullermeier. Bayes optimal multilabel classification via probabilistic classifier chains. pages 279–286, 2010. URL http://dl.acm.org/citation.cfm?id=3104322.3104359.

[33] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence in multi-label classification. Workshop proceedings of learning from multi-label data, pages 5–12, 2010.

[34] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. Machine Learning, 88(1):5–45, Jul 2012. ISSN 1573-0565. doi: 10.1007/s10994-012-5285-8. URL https://doi.org/10.1007/s10994-012-5285-8.

[35] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[36] Sahar Deppe and Volker Lohweg. Shift-invariant feature extraction for time-series motif discovery. 11 2015.

[37] Ratul Dey and Sanjay Chakraborty. Convex-hull amp; dbscan clustering to predict future weather. In 2015 International Conference and Workshop on Computing and Communication (IEMCON), pages 1–8, 2015. doi: 10.1109/IEMCON.2015.7344438.

[38] Mohamed Dilmi, Laurent Barthes, Cécile Mallet, and Aymeric Chazottes. Iterative multiscale dynamic time warping (ims-dtw): A tool for rainfall time series comparison. International Journal of Data Science and Analytics, 07 2019. doi: 10.1007/s41060-019-00193-1.

[39] Mohamed Djallel Dilmi, Laurent Barthès, Cécile Mallet, and Aymeric Chazottes. Iterative multiscale dynamic time warping (ims-dtw): a tool for rainfall time series comparison. International Journal of Data Science and Analytics, 10(1): 65–79, 2020.

[40] Brett Drury, Jorge Valverde-Rebaza, Maria Moura, and Alneu Lopes. A survey of the applications of bayesian networks in agriculture. Engineering Applications of Artificial Intelligence, 65:29 – 42, 10 2017. doi: 10.1016/j.engappai.2017.07.003.

[41] Weiyan Duan, Fanping Meng, Hongwu Cui, Yufei Lin, Guoshan Wang, and Jiangyue Wu. Ecotoxicity of phenol and cresols to aquatic organisms: a review. Ecotoxicology and Environmental Safety, 157:441–456, 2018.

[42] Sean R Eddy. Hidden markov models. Current opinion in structural biology, 6 (3):361–365, 1996.

[43] Hansen Erik. Computer aided control and monitoring of aquaculture plants. Modeling, Identification and Control, 8, 01 1987. doi: 10.4173/mic.1987.1.4.

[44] Philippe Esling and Carlos Agon. Time-series data mining. ACM Computing Surveys, 45(1):12, 2012. doi: 10.1145/2379776.2379788. URL https://hal.archives-ouvertes.fr/hal-01577883.

[45] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of 2nd International Conference on Knowledge Discovery and, pages 226–231, 1996.

[46] FAO. Food and agriculture organization of the united nations. aquaculture. 2018. URL http://www.fao.org/aquaculture/en/.

[47] Konstantinos P. Ferentinos. Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145:311–318, 2018. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2018.01.009. URL https://www.sciencedirect.com/science/article/pii/S0168169917311742.

[48] Arthur F.A. Fernandes, Eduardo M. Turra, Érika R. de Alvarenga, Tiago L. Passafaro, Fernando B. Lopes, Gabriel F.O. Alves, Vikas Singh, and Guilherme J.M. Rosa. Deep learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in nile tilapia. Computers and Electronics in Agriculture, 170: 105274, 2020. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2020.105274. URL https://www.sciencedirect.com/science/article/pii/S0168169919311561.

[49] J.G. Ferreira, A.J.S. Hawkins, and S.B. Bricker. Management of productivity, environmental effects and profitability of shellfish aquaculture the farm aquaculture resource management (farm) model. Aquaculture, 264(1):160 – 174, 2007. ISSN 0044-8486. doi: https://doi.org/10.1016/j.aquaculture.2006.12.017. URL http://www.sciencedirect.com/science/article/pii/S0044848606009094.

[50] J.G. Ferreira, L. Falconer, J. Kittiwanich, L. Ross, C. Saurel, K. Wellman, C.B. Zhu, and P. Suvanachai. Analysis of production and environmental effects of nile tilapia and white shrimp culture in thailand. Aquaculture, 447:23 – 36, 2015. ISSN 0044-8486. doi: https://doi.org/10.1016/j.aquaculture.2014.08.042. URL http://www.sciencedirect.com/science/article/pii/S0044848614004463.

[51] R.B. Fisher, Y.H. Chen-Burger, D. Giordano, L. Hardman, and F.P. Lin. Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data. Intelligent Systems Reference Library. Springer International Publishing, 2016. ISBN 9783319302065. URL https://books.google.com/books?id=j846jwEACAAJ.

[52] Rafael Garcia, Josep Quintana, Ricard Prados, Alexander Tempelaar, Nuno Gracias, Shale Rosen, Håvard Vågstøl, and Krisoffer Løvall. Automatic segmentation of fish using deep learning with application to fish size measurement. ICES Journal of Marine Science, 77, 10 2019. doi: 10.1093/icesjms/fsz186.

[53] Z. Geler, V. Kurbalija, M. Ivanović, M. Radovanović, and W. Dai. Dynamic time warping: Itakura vs sakoe-chiba. In 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), pages 1–6, 2019.

[54] Zoltan Geler, Vladimir Kurbalija, Mirjana Ivanović, Miloš Radovanović, and Weihui Dai. Dynamic time warping: Itakura vs sakoe-chiba. In 2019

IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), pages 1–6, 2019. doi: 10.1109/INISTA.2019.8778300.

[55] Gergonne. Philosophie mathématique. Considérations philosophiques sur les élémens de la science de l'étendue. Annales de mathématiques pures et appliquées, 16:209–231, 1825-1826. URL http://www.numdam.org/item/AMPA_1825-1826__16__209_0/.

[56] J. L. Giraudel, D. Aurelle, P. Berrebi, and S. Lek. Application of the Self-Organizing Mapping and Fuzzy Clustering to Microsatellite Data: How to Detect Genetic Structure in Brown Trout (Salmo trutta) Populations, pages 187–202. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000. ISBN 978-3-642-57030-8. doi: 10.1007/978-3-642-57030-8_13. URL https://doi.org/10.1007/978-3-642-57030-8_13.

[57] Rafael Giusti and Gustavo EAPA Batista. An empirical comparison of dissimilarity measures for time series classification. In 2013 Brazilian Conference on Intelligent Systems, pages 82–88. IEEE, 2013.

[58] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Å. Nielsen, and Lars Kai Hansen. On clustering fmri time series. NeuroImage, 9(3):298–310, 1999. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.1998.0391. URL https://www.sciencedirect.com/science/article/pii/S1053811998903913.

[59] Cyril Goutte, Lars Kai Hansen, Matthew G Liptrot, and Egill Rostrup. Feature-space clustering for fmri meta-analysis. Human brain mapping, 13(3):165–183, 2001.

[60] B. Guinand, K. T. Scribner, A. Topchy, K. S. Page, W. Punch, and M. K. Burnham-Curtis. Sampling issues affecting accuracy of likelihood-based classification using genetical data, pages 245–259. Springer Netherlands, Dordrecht, 2004. ISBN 978-94-007-0983-6. doi: 10.1007/978-94-007-0983-6_20. URL https://doi.org/10.1007/978-94-007-0983-6_20.

[61] Niken Gusmawati, Benoêt Soulard, Nazha Selmaoui-Folcher, Christophe Proisy, Akhmad Mustafa, Romain Le Gendre, Thierry Laugier, and Hugues Lemonnier. Surveying shrimp aquaculture pond activity using multitemporal VHSR satellite images - case study from the Perancak estuary, Bali, Indonesia. Marine Pollution Bulletin, 131(part B):49–60, 2018. doi: 10.1016/j.marpolbul.2017.03.059. URL https://hal.archives-ouvertes.fr/hal-01558070.

[62] Niken Gusmawati, Benoît Soulard, Nazha Selmaoui-Folcher, Christophe Proisy, Akhmad Mustafa, Romain Le Gendre, Thierry Laugier, and Hugues Lemonnier. Surveying shrimp aquaculture pond activity using multitemporal vhsr satellite images - case study from the perancak estuary, bali, indonesia. Marine Pollution Bulletin, 131:49 – 60, 2018. ISSN 0025-326X. doi: https://doi.org/10.1016/j.marpolbul.2017.03.059. URL http://www.sciencedirect.com/science/article/pii/S0025326X17302795. Special Issue: Indonesia seas management.

[63] Ben F Hajek and Claude E Boyd. Rating soil and water information for aquaculture. Aquacultural engineering, 13(2):115–128, 1994.

[64] Jichong Han, Zhao Zhang, Juan Cao, Yuchuan Luo, Liangliang Zhang, Ziyue Li, and Jing Zhang. Prediction of winter wheat yield based on multi-source data and machine learning in china. Remote Sensing, 12(2), 2020. ISSN 2072-4292. doi: 10.3390/rs12020236. URL https://www.mdpi.com/2072-4292/12/2/236.

[65] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. JSTOR: Applied Statistics, 28(1):100–108, 1979.

[66] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

[67] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A, 79(8):2554–2558, April 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC346238/.

[68] J.J. Hopfield. Artificial neural networks. IEEE Circuits and Devices Magazine, 4(5):3–10, 1988. doi: 10.1109/101.8118.

[69] Juan Huan, Hui Li, Mingbao Li, and Bo Chen. Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of chang zhou fishery demonstration base, china. Computers and Electronics in Agriculture, 175:105530, 08 2020. doi: 10.1016/j.compag.2020.105530.

[70] Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, and Shaokai Wang. Time series k-means: A new k-means type smooth subspace clustering for time series data. Information Sciences, 367-368:1 – 13, 2016. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2016.05.

040. URL http://www.sciencedirect.com/science/article/pii/S0020025516303796.

[71] Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond YK Lau, Nan Jiang, and Shaokai Wang. Time series k-means: A new k-means type smooth subspace clustering for time series data. Information Sciences, 367:1–13, 2016.

[72] J Stuart Hunter. The exponentially weighted moving average. Journal of quality technology, 18(4):203–210, 1986.

[73] F. Itakura. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23(1):67–72, 1975.

[74] CJ Jackson and Y-G Wang. Modelling growth rate of penaeus monodon fabricius in intensively managed ponds: effects of temperature, pond age and stocking density. Aquaculture research, 29(1):27–36, 1998.

[75] Ahsan Jalal, Ahmad Salman, Ajmal Mian, Mark Shortis, and Faisal Shafait. Fish detection and species classification in underwater environments using deep learning with temporal information. Ecological Informatics, 57:101088, 04 2020. doi: 10.1016/j.ecoinf.2020.101088.

[76] Goncalves Jardim and Rosa Ricardo, Luis. Aquaculture production optimization through enhanced data analytics. 2016. Sem PDF 6th Offshore Mariculture Conference.

[77] Elsa Jesus, Andreia Artifice, JoÃ£o Sarraipa, Gary Mcmanus, and Fernando Luis-Ferreira. A training programme to support aquasmart project exploitation. 07 2018.

[78] Leilei Jin and Hong Liang. Deep learning for underwater image recognition in small sample size situations. In OCEANS 2017-Aberdeen, pages 1–4. IEEE, 2017.

[79] Joao and Rihtar. Data analytics in aquaculture. SIKDD 2016, 2016.

[80] Rihtar Joao, Kostas Sarraipa, and Victor Seferis. Data analytics : models and algorithms for intelligent data analysis. 2016.

[81] S. C. Johnson. Hierarchical clustering schemes. Psychometrika, 2:241–254, 1967. URL www.garfield.library.upenn.edu/classics1985/A1985AQU6100001.pdf.

[82] K. Kalpakis, D. Gada, and V. Puttagunta. Distance measures for effective clustering of arima time-series. In ICDM, pages 273–280, 2001.

[83] Konstantinos Kalpakis, Dhiral Gada, and Vasundhara Puttagunta. Distance measures for effective clustering of arima time-series. In Proceedings 2001 IEEE international conference on data mining, pages 273–280. IEEE, 2001.

[84] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. IEEE access, 6:1662–1669, 2017.

[85] Dhian Satria Yudha Kartika and Darlis Herumurti. Koi fish classification based on hsv color space. In 2016 International Conference on Information Communication Technology and Systems (ICTS), pages 96–100, 2016. doi: 10.1109/ICTS.2016.7910280.

[86] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 173:24–49, 2021.

[87] Noor Kamal Kaur, Usvir Kaur, and Dheerendra Singh. K-medoid clustering algorithm-a review. Int. J. Comput. Appl. Technol, 1(1):42–45, 2014.

[88] Oscar Kempthorne. The correlation between relatives on the supposition of mendelian inheritance. American journal of human genetics, 20(4):402, 1968.

[89] Eamonn Keogh. Efficiently finding arbitrarily scaled patterns in massive time series databases. In PKDD 2003, pages 253–265. Springer Berlin Heidelberg, 2003.

[90] Eamonn Keogh and Michael Pazzani. Derivative dynamic time warping. First SIAM International Conference on Data Mining, 1, 01 2002. doi: 10.1137/1.9781611972719.1.

[91] Eamonn Keogh, Li Wei, Xiaopeng Xi, Michalis Vlachos, Sang-Hee Lee, and Pavlos Protopapas. Supporting exact indexing of arbitrarily rotated shapes and periodic time series under euclidean and warping distance measures. VLDB J., 18:611–630, 06 2009. doi: 10.1007/s00778-008-0111-4.

[92] Alison King, Zeb Tonkin, and Jason Lieschke. Short-term effects of a prolonged blackwater event on aquatic fauna in the murray river, australia: Considerations for future events. Marine and Freshwater Research, 63:576–586, 06 2012. doi: 10.1071/MF11275.

[93] T. Kohonen. the self-organizing map. Neurocomputing, 21:1, 1998.

[94] Zhi Hong Kok, Abdul Rashid Mohamed Shariff, Meftah Salem M. Alfatni, and Siti Khairunniza-Bejo. Support vector machine in precision agriculture: A review. Computers and Electronics in Agriculture, 191: 106546, 2021. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2021.106546. URL https://www.sciencedirect.com/science/article/pii/S0168169921005639.

[95] B. Kosko. Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence. 1991.

[96] Thales Sehn Körting, 2014. URL http://www.dpi.inpe.br/~tkorting/.

[97] Gilles Le Moullac and Philippe Haffner. Environmental factors affecting immune responses in crustacea. Aquaculture, 191(1-3):121–131, 2000.

[98] Gilles Le Moullac, Claude Soyez, Denis Saulnier, Dominique Ansquer, Jean Christophe Avarre, and Peva Levy. Effect of hypoxic stress on the immune response and the resistance to vibriosis of the shrimppenaeus stylirostris. Fish and Shellfish Immunology, 8(8):621–629, 1998.

[99] Phillip G. Lee. A review of automated control systems for aquaculture and design criteria for their implementation. Aquacultural Engineering, 14(3):205–227, 1995. ISSN 0144-8609. doi: https://doi.org/10.1016/0144-8609(94)00002-I. URL https://www.sciencedirect.com/science/article/pii/014486099400002I.

[100] Phillip G Lee. Process control and artificial intelligence software for aquaculture. Aquacultural Engineering, 23(1):13 – 36, 2000. ISSN 0144-8609. doi: https://doi.org/10.1016/S0144-8609(00)00044-3. URL http://www.sciencedirect.com/science/article/pii/S0144860900000443.

[101] Pierrette Lemaire, E Bernard, JA Martinez-Paz, and Liet Chim. Combined effect of temperature and salinity on osmoregulation of juvenile and subadult penaeus stylirostris. Aquaculture, 209(1-4):307–317, 2002.

[102] Hugues Lemonnier and Sébastien Faninoz. Effect of water exchange on effluent and sediment characteristics and on partial nitrogen budget in semi-intensive shrimp ponds in new caledonia. Aquaculture research, 37(9):938–948, 2006.

[103] Hugues Lemonnier, Eric Bernard, Eric Boglio, Cyrille Goarant, and Jean-Claude Cochard. Influence of sediment characteristics on shrimp physiology: ph as principal effect. Aquaculture, 240(1-4):297–312, 2004.

[104] Hugues Lemonnier, Alain Herbland, Lucas Salery, and Benoît Soulard. "summer syndrome" in litopenaeus stylirostris grow out ponds in new caledonia: zootechnical and environmental factors. Aquaculture, 261(3):1039–1047, 2006.

[105] Hugues Lemonnier, Florence Royer, Florian Caradec, Etienne Lopez, Clarisse Hubert, Émilie Rabiller, Térence Desclaux, Jean-Michel Fernandez, and Françoise Andrieux-Loyer. Diagenetic processes in aquaculture ponds showing metal accumulation on shrimp gills. Frontiers in Marine Science, 8:625789, 2021.

[106] Hugues Lemonnier, Nelly Wabete, Dominique Pham, Jean-Hervé Lignot, Kiam Barri, Isabelle Mermoud, Florence Royer, Viviane Boulo, and Thierry Laugier. Iron deposits turn blue shrimp gills to orange. Aquaculture, 540:736697, 2021.

[107] Hailin Li. Multivariate time series clustering based on common principal component analysis. Neurocomputing, 349:239–247, 2019.

[108] Xiu Li, Min Shang, Jing Hao, and Zhixiong Yang. Accelerating fish detection and recognition by sharing cnns with objectness learning. In OCEANS 2016-Shanghai, pages 1–5. IEEE, 2016.

[109] Yang Li, Guowei Wang, Yu Chen, Yang Jiao, Haijiao Yu, and Guogang Zhao. Application of DBSCAN Algorithm in Precision Fertilization Decision of Maize. In Daoliang Li and Chunjiang Zhao, editors, 11th International Conference on Computer and Computing Technologies in Agriculture (CCTA volume AICT-546 of Computer and Computing Technologies in Agriculture XI, pages 453–459, Jilin, China, August 2017. Springer International Publishing. doi: 10.1007/978-3-030-06179-1_45. URL https://hal.inria.fr/hal-02111561.

[110] Konstantinos Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. Sensors, 18:2674, 08 2018. doi: 10.3390/s18082674.

[111] J-H Lignot, JC Cochard, C Soyez, P Lemaire, and G Charmantier. Osmoregulatory capacity according to nutritional status, molt stage and body weight in penaeus stylirostris. Aquaculture, 170(1):79–92, 1999.

[112] Kai Lorenzen. Toward a new paradigm for growth modeling in fisheries stock assessments: Embracing plasticity and its consequences. Fisheries Research, 180, 01 2016. doi: 10.1016/j.fishres.2016.01.006.

[113] JingGui Lu, Yi Liu, and Xiaoli Li. The decision tree application in agricultural development. In Hepu Deng, Duoqian Miao, Jingsheng Lei, and Fu Lee Wang, editors, Artificial Intelligence and Computational Intelligence, pages 372–379, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23881-9.

[114] Oscar Luaces, Jorge Díez, Jose Barranquero, Juan del Coz, and Antonio Bahamonde. Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence, 1, 12 2012. doi: 10.1007/s13748-012-0030-x.

[115] Jacqueline Léon. Ségal, jérôme, le zéro et le un : histoire de la notion scientifique d'information au 20e siècle (2003). Histoire Épistémologie Langage, 29(1):171–173, 2007. URL https://www.persee.fr/doc/hel_0750-8069_2007_num_29_1_2918_t10_0171_0000_1. Included in a thematic issue : Histoire des théories du son.

[116] Elizabeth Ann Maharaj. Cluster of time series. Journal of Classification, 17(2), 2000.

[117] Michael C. Melnychuk, Emily Peterson, Matthew Elliott, and Ray Hilborn. Fisheries management impacts on target species status. Proceedings of the National Academy of Sciences, 114(1):178–183, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1609915114. URL https://www.pnas.org/content/114/1/178.

[118] Tom Mitchell, Bruce Buchanan, Gerald DeJong, Thomas Dietterich, Paul Rosenbloom, and Alex Waibel. Machine learning. Annual review of computer science, 4(1):417–433, 1990.

[119] Tom M Mitchell and Tom M Mitchell. Machine learning, volume 1. McGraw-hill New York, 1997.

[120] Graham Monkman, Kieran Hyder, Michel Kaiser, Franck Vidal, G Monkman, and K Kaiser. Accurate estimation of fish length in single camera photogrammetry with a fiducial marker. ICES Journal of Marine Science, 77, 02 2019. doi: 10.1093/icesjms/fsz030.

[121] Vania C. Mota, Flavio A. Damasceno, and Daniel F. Leite. Fuzzy clustering and fuzzy validity measures for knowledge discovery and decision making in

agricultural engineering. Computers and Electronics in Agriculture, 150:118–124, 2018. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2018.04.011. URL https://www.sciencedirect.com/science/article/pii/S0168169918300632.

[122] Chantal Mugnier and Claude Soyez. Response of the blue shrimp litopenaeus stylirostris to temperature decrease and hypoxia in relation to molt stage. Aquaculture, 244(1-4):315–322, 2005.

[123] Sneha Murmu and Sujata Biswas. Application of fuzzy logic and neural network in crop classification: A review. Aquatic Procedia, 4:1203–1210, 2015. ISSN 2214-241X. doi: https://doi.org/10.1016/j.aqpro.2015.02.153. URL https://www.sciencedirect.com/science/article/pii/S2214241X15001546. INTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE'15).

[124] Meinard Müller. Dynamic Time Warping, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74048-3. doi: 10.1007/978-3-540-74048-3_4. URL https://doi.org/10.1007/978-3-540-74048-3_4.

[125] Xanthoula Pantazi, Dimitrios Moshou, Roberto Oberti, Jon West, Abdul Mouazen, and Dionysios Bochtis. Detection of biotic and abiotic stresses in crops by using hierarchical self organizing classifiers. Precision Agriculture, 18:1–11, 06 2017. doi: 10.1007/s11119-017-9507-8.

[126] X.E. Pantazi, Afroditi Alexandra Tamouridou, Thomas Alexandridis, Anastasia Lagopodi, G. Kontouris, and Dimitrios Moshou. Detection of silybum marianum infection with microbotryum silybum using vnir field spectroscopy. Computers and Electronics in Agriculture, 137:130–137, 05 2017. doi: 10.1016/j.compag.2017.03.017.

[127] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. ACM SIGMOD Record, 45:69–76, 06 2016. doi: 10.1145/2949741.2949758.

[128] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. ACM SIGMOD Record, 45:69–76, 06 2016. doi: 10.1145/2949741.2949758.

[129] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In Icml, volume 1, pages 727–734, 2000.

[130] Ricardus Anggi Pramunendar, Sunu Wibirama, and Paulus Insap Santosa. Fish classification based on underwater image interpolation and back-propagation neural network. In 2019 5th International Conference on Science and Technology (ICST), volume 1, pages 1–6, 2019. doi: 10.1109/ICST47872.2019.9166295.

[131] Hongwei Qin, Xiu Li, Jian Liang, Yigang Peng, and Changshui Zhang. Deepfish: Accurate underwater live fish recognition with a deep architecture. Neurocomputing, 187:49–58, 2016. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.10.122. URL https://www.sciencedirect.com/science/article/pii/S0925231215017312. Recent Developments on Deep Big Vision.

[132] Maurice Henry Quenouille. A large-sample test for the goodness of fit of autoregressive schemes. Journal of the Royal Statistical Society, 110:123–129, 1947.

[133] J. R. Quinlan. Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20(2):339–346, 1990. doi: 10.1109/21.52545.

[134] LR Rabiner and JG Wilpon. Considerations in applying clustering techniques to speaker-independent word recognition. The Journal of the Acoustical Society of America, 66(3):663—673, September 1979. ISSN 0001-4966. doi: 10.1121/1.383693. URL https://doi.org/10.1121/1.383693.

[135] Muhammad Naufal Rachmatullah and Iping Supriana. Low resolution image fish classification using convolutional neural network. In 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA), pages 78–83, 2018. doi: 10.1109/ICAICTA.2018.8541313.

[136] Ashfaqur Rahman and MD Shahriar. Algae growth prediction through identification of influential environmental variables: A machine learning approach. International Journal of Computational Intelligence and Applications, 12, 06 2013. doi: 10.1142/S1469026813500089.

[137] Yannick Ramage, Benoit Soulard, Benoit Beliaeff, and Jose Herlin. Détermination d'indicateurs de performance des élevages de crevettes en nouvelle-calédonie. 2011.

[138] Marco Ramoni, Paola Sebastiani, and Paul Cohen. Bayesian clustering by dynamics. Machine learning, 47(1):91–121, 2002.

[139] PJ Ramos, Flavio Augusto Prieto, EC Montoya, and Carlos Eugenio Oliveros. Automatic fruit count on coffee branches using computer vision. Computers and Electronics in Agriculture, 137:9–22, 2017.

[140] LAWANI Ran and A Djikpo. Effets des pratiques agricoles sur la pollution des eaux de surface en république du bénin. Larhyss Journal, 30:2017, 2017.

[141] N. H. Rao, P. B. S. Sarma, and Subhash Chander. Real-time adaptive irrigation scheduling under a limited water supply. Agricultural Water Management, 20 (4):267–279, February 1992. URL https://ideas.repec.org/a/eee/agiwat/v20y1992i4p267-279.html.

[142] Lars Ravensbeck, Ayoe Hoff, and Hans Frost. Implications for fisheries management by inclusion of marine ecosystem services. IFRO Working Paper 2016/12, Copenhagen, 2016. URL http://hdl.handle.net/10419/204404.

[143] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In Joint European conference on machine learning and knowledge discovery in databases, pages 254–269. Springer, 2009.

[144] Irina Rish. An empirical study of the naive bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3, pages 41–46. IBM New York, 2001.

[145] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. pages 410–420, 01 2007.

[146] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 410–420, 2007.

[147] Georg Ruß and Rudolf Kruse. Exploratory hierarchical clustering for management zone delineation in precision agriculture. In Petra Perner, editor, Advances in Data Mining. Applications and Theoretical Aspects, pages 161–173, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23184-1.

[148] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978.

[149] Ahmad Salman, Ahsan Jalal, Faisal Shafait, Ajmal Mian, Mark Shortis, James Seager, and Euan Harvey. Fish species classification in unconstrained underwater environments based on deep learning. Limnology and Oceanography: Methods, 14(9):570–585, 2016.

[150] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. volume 11, pages 70–80, 01 2004.

[151] Lawrence Saul and Michael Jordan. Boltzmann chains and hidden markov models. In G. Tesauro, D. Touretzky, and T. Leen, editors, Advances in Neural Information Processing Systems, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper/1994/file/4e0cb6fb5fb446d1c92ede2ed8780188-Paper.pdf.

[152] Subhajit Sengupta and Won Suk Lee. Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. Biosystems Engineering, 117:51–61, 2014.

[153] S Shalini, J. Muruganandham, and R Jayasri. Nutrition deficiency detection in leaves using k-means clustering algorithm. Journal of Physics: Conference Series, 1717:012003, 01 2021. doi: 10.1088/1742-6596/1717/1/012003.

[154] Robert H Shumway and David S Stoffer. Arima models. In Time series analysis and its applications, pages 75–163. Springer, 2017.

[155] Ristic Mirjana Dj— Peric-Grujic Aleksandra A—0000-0002-2593-4796 Pocajt Viktor V— Siljic-Tomic Aleksandra—, Antanasijevic Davor Z—0000-0002-0915-1281. A linear and non-linear polynomial neural network modeling of dissolved oxygen content in surface water: Inter- and extrapolation performance with inputs' significance analysis. SCIENCE OF THE TOTAL ENVIRONMENT, 610:1038–1046, 2018.

[156] Benoit Soulard, Julie Frappier, Jose Herlin, and Beliaeff Benoit. Stylog : base de données pour le suivi des élevages de crevettes de nouvelle-calédonie, 2009. URL https://archimer.ifremer.fr/doc/00065/17659/.

[157] D. Steinley. Properties of the hubert-arabie adjusted rand index. Psychological methods, 9 3:386–96, 2004.

[158] Adrian Stetco, Xiao-Jun Zeng, and John A. Keane. Fuzzy c-means++: Fuzzy c-means with effective seeding initialization. Expert Syst. Appl., 42(21): 7541–7548, 2015. URL http://dblp.uni-trier.de/db/journals/eswa/eswa42.html#StetcoZK15.

[159] Zbigniew R. Struzik and Arno Siebes. The haar wavelet transform in the time series similarity paradigm. In Jan M. Żytkow and Jan Rauch, editors, Principles of Data Mining and Knowledge Discovery, pages 12–22, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48247-5.

[160] Xin Sun, Junyu Shi, Junyu Dong, and Xinhua Wang. Fish recognition from low-resolution underwater images. In 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 471–476, 2016. doi: 10.1109/CISP-BMEI.2016.7852757.

[161] P. Surya and Laurence Aroquiaraj. Performance analysis of k-means and k-medoid clustering algorithms using agriculture dataset. Journal of Emerging Technologies and Innovative Research (JETIR), 6, 2019. URL https://ssrn.com/abstract=3345800.

[162] S Syahminan, J Maknunah, R Dijaya, and H Hindarto. KNN (k-nearby neighbor) for identifying agricultural land. Journal of Physics: Conference Series, 1402 (6):066059, dec 2019. doi: 10.1088/1742-6596/1402/6/066059. URL https://doi.org/10.1088/1742-6596/1402/6/066059.

[163] Alaa Tharwat, Ahmed Abdelmonem Hemedan, Aboul Ella Hassanien, and Thomas Gabel. A biometric-based model for fish species classification. Fisheries Research, 204:324–336, 2018. ISSN 0165-7836. doi: https://doi.org/10.1016/j.fishres.2018.03.008. URL https://www.sciencedirect.com/science/article/pii/S0165783618300821.

[164] Xijun Tian, PingSun Leung, and Eithan Hochman. Shrimp growth functions and their economic implications. Aquacultural Engineering, 12(2):81 – 96, 1993. ISSN 0144-8609. doi: https://doi.org/10.1016/0144-8609(93)90018-7. URL http://www.sciencedirect.com/science/article/pii/0144860993900187.

[165] K. Tjorve and E. Tjorve. The use of gompertz models in growth analyses, and new gompertz-model approach: An addition to the unified-richards family. PLOS ONE, 12(6):1–17, 06 2017. doi: 10.1371/journal.pone.0178691. URL https://doi.org/10.1371/journal.pone.0178691.

[166] Jannaï Tokotoko, , Hugues Lemonnier, and Nazha Selmaoui-Folcher. 2020. URL https://gt-gast.irisa.fr/programme-gast-2020/.

[167] Jannaï Tokotoko, Nazha Selmaoui-Folcher, Rodrigue Govan, and Hugues Lemonnier. Tsx-means: An optimal K search approach for time series clustering. In Christine Strauss, Gabriele Kotsis, A Min Tjoa, and Ismail Khalil, editors, Database and Expert Systems Applications - 32nd International Conference, DEXA 2021, Virtual Event, September 27-30, 2021, Proceedings, Part II, volume 12924 of Lecture Notes in Computer Science, pages 232–238. Springer, 2021. doi: 10.1007/978-3-030-86475-0_23. URL https://doi.org/10.1007/978-3-030-86475-0_23.

[168] Jannaï Tokotoko, Rodrigue Govan, Hugues Lemonnier, and Nazha Selmaoui-Folcher. Multiscale and multivariate time series clustering: A new approach. In Michelangelo Ceci, Sergio Flesca, Elio Masciari, Giuseppe Manco, and Zbigniew W. Ras, editors, Foundations of Intelligent Systems - 26th International Symposium, ISMIS 2022, Cosenza, Italy, October 3-5, 2022, Proceedings, volume 13515 of Lecture Notes in Computer Science, pages 283–293. Springer, 2022. doi: 10.1007/978-3-031-16564-1_27. URL https://doi.org/10.1007/978-3-031-16564-1_27.

[169] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3:1–13, 09 2009. doi: 10.4018/jdwm.2007070101.

[170] Grigorios Tzortzis and Aristidis Likas. The global kernel k-means clustering algorithm. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pages 1977–1984, 2008. doi: 10.1109/IJCNN.2008.4634069.

[171] J. Ulmo. Différents aspects de l'analyse discriminante. Revue de Statistique Appliquée, 21(2):17–55, 1973. URL http://www.numdam.org/item/RSA_1973__21_2_17_0/.

[172] Vladimir N. Vapnik. Statistical Learning Theory. Wiley-Interscience, September 1998. ISBN 0471030031. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471030031.

[173] M. Vlachos, G. Kollios, and D. Gunopulos. Discovering similar multidimensional trajectories. In ICDE'02, pages 673–684, 2002.

[174] Michalis Vlachos, Jessica Lin, Eamonn Keogh, and Dimitrios Gunopulos. A wavelet-based anytime algorithm for k-means clustering of time series. Proc. Workshop on Clustering High Dimensionality Data and its Applications, 04 2003.

[175] Nelly Wabete, Liet Chim, Dominique Pham, Pierrette Lemaire, and Jean-Charles Massabuau. A soft technology to improve survival and reproductive performance of litopenaeus stylirostris by counterbalancing physiological disturbances associated with handling stress. Aquaculture, 260(1-4):181–193, 2006.

[176] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey, 2019. URL https://arxiv.org/abs/1902.06068.

[177] J. Wilpon and L. Rabiner. A modified k-means clustering algorithm for use in isolated work recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(3):587–594, 1985. doi: 10.1109/TASSP.1985.1164581.

[178] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.

[179] Jun Wu, Anastasiya olesnikova, Chi-Hwa Song, and Won Lee. The development and application of decision tree for agriculture data. pages 16 – 20, 02 2009. doi: 10.1109/IITSI.2009.10.

[180] Natalia Yakovleva, Joseph Sarkis, and Thomas Sloan. Sustainable benchmarking of supply chains: the case of the food industry. International Journal of Production Research, 50(5):1297–1317, 2012. doi: 10.1080/00207543.2011.571926. URL https://doi.org/10.1080/00207543.2011.571926.

[181] Shlomo Yitzhaki. Stochastic dominance, mean variance, and gini's mean difference. American Economic Review, 72:178–85, 1982.

[182] Run Yu, PingSun Leung, and Paul Bienfang. Predicting shrimp growth: Artificial neural network versus nonlinear regression models. Aquacultural Engineering, 34(1):26 – 32, 2006. ISSN 0144-8609. doi: https://doi.org/10.1016/j.aquaeng.2005.03.003. URL http://www.sciencedirect.com/science/article/pii/S0144860905000348.

[183] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition, 40(7):2038–2048, 2007.

[184] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. Knowledge and Data Engineering, IEEE Transactions on, 26:1819–1837, 08 2014. doi: 10.1109/TKDE.2013.39.

[185] Yifan Zhang, Peter Fitch, and Peter Thorburn. Predicting the trend of dissolved oxygen based on the kpca-rnn model. Water, 12:585, 02 2020. doi: 10.3390/w12020585.

[186] Shili Zhao, Song Zhang, Jincun Liu, He Wang, Jia Zhu, Daoliang Li, and Ran Zhao. Application of machine learning in intelligent fish aquaculture: A review. Aquaculture, 540:736724, 2021. ISSN 0044-8486. doi: https://doi.org/10.1016/j.aquaculture.2021.736724. URL https://www.sciencedirect.com/science/article/pii/S0044848621003860.

[187] CHENG Zhi. Mining recurrent patterns in a dynamic attributed graph. Application to aquaculture Pond Monitoring by satellite images. PhD thesis, University of New Caledonia, 2018.

[188] Pei-Yuan Zhou and Keith CC Chan. A model-based multivariate time series clustering algorithm. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 805–817. Springer, 2014.

LISTE DES PUBLICATIONS

1. **Jannaï Tokotoko**, Rodrigue Govan, Hugues Lemonnier, Nazha Selmaoui-Folcher: Multiscale and Multivariate Time Series Clustering: A New Approach. ISMIS 2022: 283-293

2. Nazha Selmaoui-Folcher, **Jannaï Tokotoko**, Samuel Gorohouna, Laisa Roi, Claire Leschi, Catherine Ris: Concept of Temporal Pretopology for the Analysis for Structural Changes: Application to Econometrics. Int. J. Data Warehous. Min. 18(2): 1-17 (2022)

3. **Jannaï Tokotoko**, Nazha Selmaoui-Folcher, Rodrigue Govan, Hugues Lemonnier: TSX-Means: An Optimal K Search Approach for Time Series Clustering. DEXA (2) 2021: 232-238

4. Nazha Selmaoui-Folcher, **Jannaï Tokotoko**, Samuel Gorohouna, Laisa Roi: Concept de prétopologie temporelle pour l'analyse des évolutions structurelles. EGC 2021: 119-131

5. **Jannaï Tokotoko**, Romane Scherrer, Hugues Lemonnier, Nazha Selmaoui-Folcher: *Analyse de la performance de filières aquacoles à partir de données spatio-temporelles*. Workshops 2020 Gestion et Analyse des données Spatiales et Temporelles à EGC, Bruxelles.

6. **Jannaï Tokotoko**, Frédéric Flouvat, Claire Goiran, Laetitia Hédouin, Antoine Collin, Nazha Selmaoui-Folcher: *Supervised Classification of Satellite Images with Spatially Inaccurate Training Field Data*. ICDM Workshops 2018, pp.1381-1388, Singapour.

7. **Jannaï Tokotoko**, Frédéric Flouvat, Claire Goiran, Laetitia Hédouin, Antoine Collin, Nazha Selmaoui-Folcher: *Prétraitement de données spatialement imprécises*

pour une classification supervisée basée sur les images satellitaires. Actes des 18ème Conférence Internationale francophone sur l'Extraction et Gestion des Connaissances (EGC'18), Hermann-Editions, RNTI, Vol. E-34, pp.167-1782018), Paris France. (article long).

Annexes

Annexe A

Distribution de la croissance initiale (*C*), de la vitesse de convergence et de la survie en fonction des pairs de clusters de températures hebdomadaires

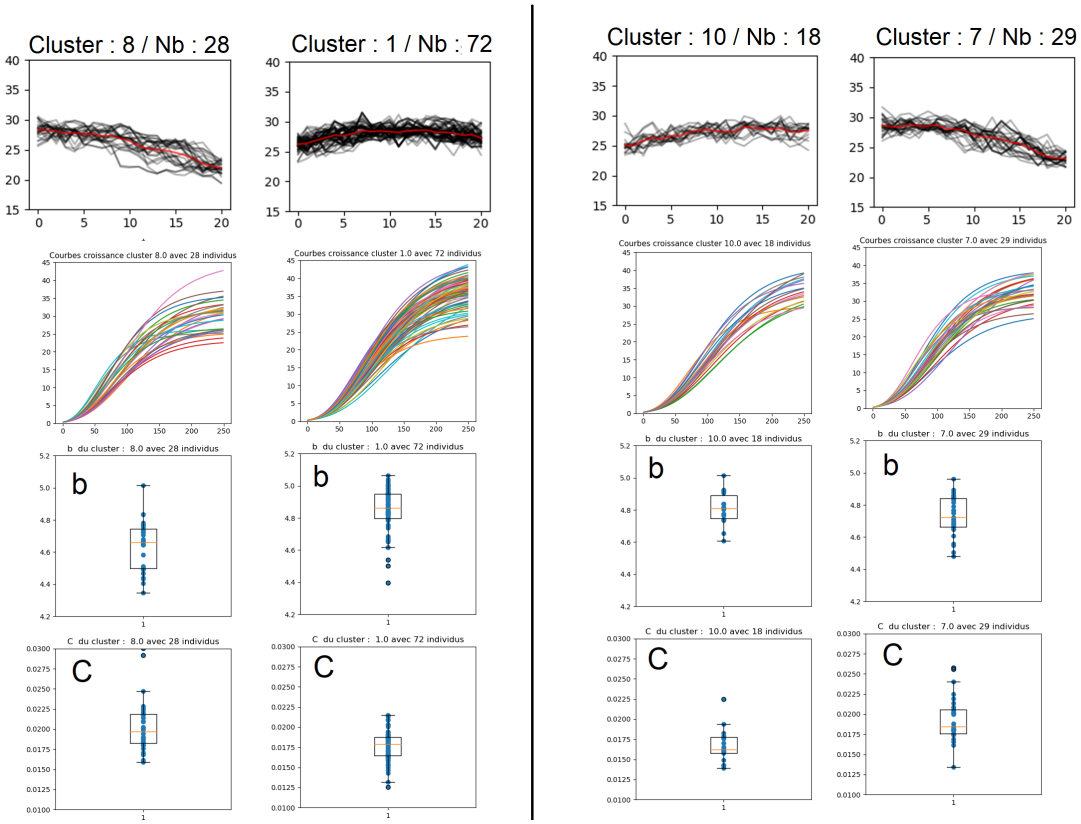

Fig. A.1 Distribution de la croissance initiale (*C*), de la vitesse de convergence et de la survie en fonction des pairs de clusters de températures hebdomadaires, avec des *p-valeurs* inférieurs à 0.05% avec la méthode *Xmeans-TS*.

Annexe B

Distribution des paramètres de Gomprertz

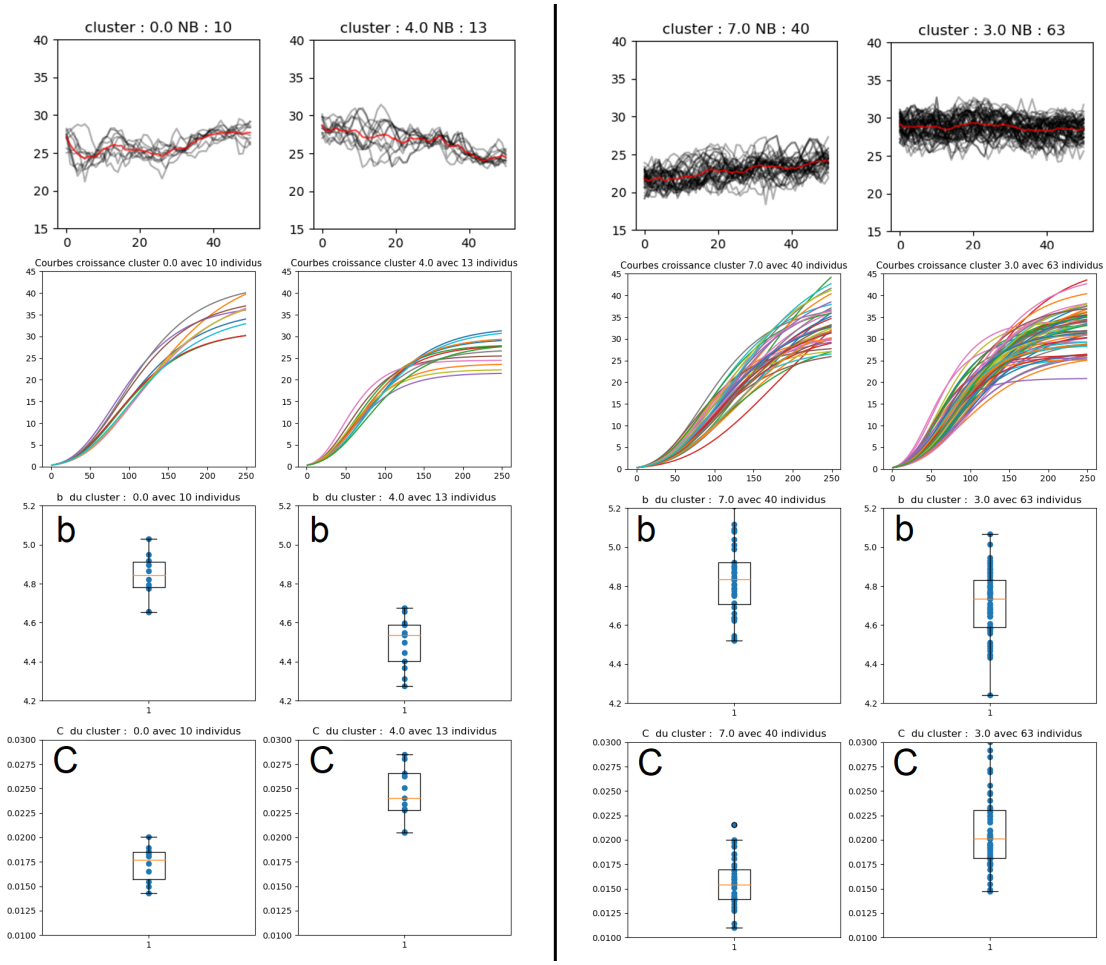

Fig. B.1 Distribution des paramètres de Gomprertz en fonction des pairs clusters de température journalière ayant les p-valeurs inférieur à 0.05 pour *Xmeans-TS*.

Annexe C

Comparaison de clusters générés par X-meansTS et K-Shape

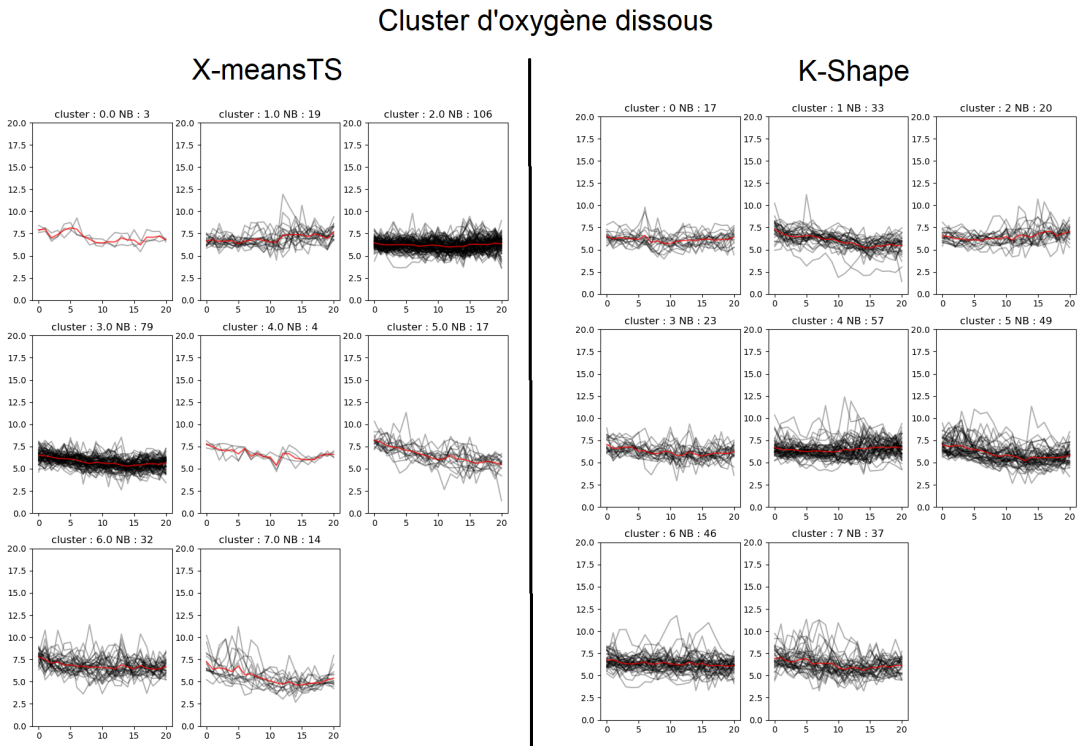

Fig. C.1 Comparaison de clusters d'oxygène dissous générés par X-meansTS et K-Shape

Annexe D

Description de Clusters multivariée multi-échelle avec des distribution des paramètres zootechniques significatives

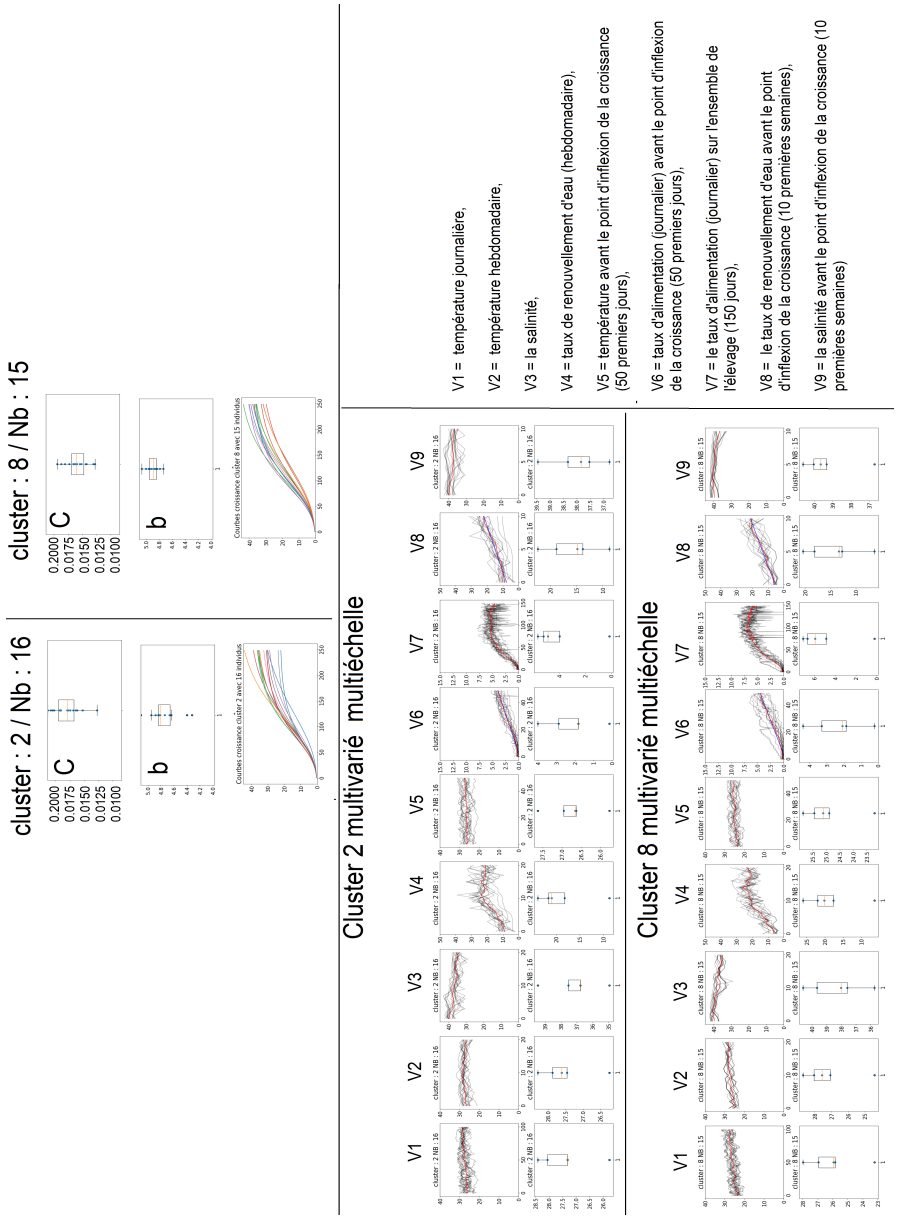

Fig. D.1 Clustering multivariée multi-échelle