Analyse de la performance de filières aquacoles à partir de données spatio-temporelles

Jannaï Tokotoko*, Romane Scherrer*, Hugues Lemonnier **, Nazha Selmaoui-Folcher *

*ISEA - Université de la Nouvelle-Calédonie, BP R4, 98851, Nouméa, Nouvelle-Calédonie
**IFREMER, 101 rue promenade Laroque, Nouméa, Nouvelle-Calédonie

Résumé. 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 cet article, nous décrirons la démarche appliquée pour analyser un jeu de données généré par de multiples acteurs et établi sur des élevages de la crevette tropicale *Litopenaeus stylirostris* réalisés en Nouvelle-Calédonie entre 2003 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 et ensuite (ii) les prédire. Vu la complexité des données, ces modèles seront appliqués, pour une étude préliminaire, à des paramètres construits, par exemple à partir d'un modèle de croissance des animaux établi 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.

1 Introduction

La multiplication des données générées dans les procédés industriels à destination de l'alimentation soulève un grand nombre de défis en matière d'analyse de données. Ces défis s'imposent de part la complexité, l'hétérogénéité et l'imprécision des données collectées à l'échelle des filières. Face à cela, la science de données vise à apporter des solutions pour analyser ces données hétérogènes.

Les systèmes intelligents intégrant l'apprentissage automatique à partir des données et permettant la gestion automatique des processus devraient assurer une meilleure compréhension des techniques de production, de l'impact du marché ainsi qu'une meilleure maîtrise des maladies et une diminution des intrants (Lee, 2000; Joao et al., 2016). L'aquaculture, industrie majeure pour les produits de la mer (FAO, 2016, 2018), est d'après la littérature, un domaine idéal pour l'application de ces méthodes (fouille de données, big data, etc.) (Joao et Rihtar, 2016).

Les travaux impliquant la science de données (apprentissage automatique, régression, optimisation, etc.) ont été appliqués dans différents domaines : la génétique (Guinand et al., 2004; Giraudel et al., 2000), l'analyse de l'activité aquacole par satellite (Gusmawati et al., 2018; Cheng et al., 2017; Zhi, 2018), l'impact de l'environnement d'élevage sur la production (Ferreira et al., 2015; Lee, 2000; Czogaa et Rawlik, 1989; Bourke et al., 1993; Jardim et Ricardo, 2016; Jesus et al., 2018) et les études sur la croissance (Charles, 1979; Yu et al., 2006; Tian et al., 1993; Rahman et Shahriar, 2013).

A notre connaissance, il n'existe toutefois pas de démarche intégrative permettant une analyse globale du fonctionnement et des résultats de ces filières, alors que de nombreux facteurs qu'ils soient zootechniques, environnementaux, sociaux, économiques sont à l'origine des performances de productions.

L'objectif de notre travail est de mettre en place une méthodologie générale pour établir un lien entre pratiques et performances des élevages. Par performances, nous associons quantité et qualité produite.

2 Description des données

Nous étudierons les données d'élevages de la filière crevetticole Calédonienne gérées par le GFA (Groupement des Fermes Aquacoles). L'étude est réalisée sur les données de 18 fermes positionnées géographiquement sur la côte ouest et réparties sur plus de 400 km du nord au sud de l'île.

Elle concerne 400 élevages de crevettes tropicales *Litopenaeus stylirostris* réalisés entre 2004 et 2014. Un élevage s'effectue sur une durée de 5 à 7 mois. Afin d'assurer un suivi à l'échelle de la filière, différents types de données sont recueillies au cours de chaque élevage (paramètres physico-chimiques, poids moyens...).

2.1 Les données d'élevage

Les données d'élevages proviennent de la base de données appelée STYLIBASE (Soulard et al. (2009)) qui est une base de données relationnelles normalisée (3FN). Elle a pour vocation de rassembler au sein d'une même base fonctionnelle, les données issues des bases unitaires alimentées indépendamment par chaque ferme Calédonienne.

Ces données représentent des séries temporelles associées aux évolutions des indicateurs de qualité du milieu (température, oxygène dissous dans l'eau..) et des indicateurs zootechniques (évolution du poids moyen, densité initiale de crevettes...) relevées au cours de chaque élevage.

L'ordre de grandeur de la quantité de données associées uniquement aux suivis temporels de qualité du milieu est d'environ 100000 données par an.

2.2 Les données de qualité du produit

La qualité du produit issu des élevages est évaluée par la SOPAC (Société des Producteurs Aquacoles Calédoniens) à partir d'une analyse qualitative et quantitative en laboratoire et est déterminée selon les différents calibres des crevettes pêchées. Ces évaluations sont effectuées sur 3 prélèvements (dites pêches) espacés d'environ un mois à la fin d'un élevage. La qualité est définie selon une estimation de la présence de défauts dans la production. La figure 1 montre l'évolution annuelle entre 2005 et 2013 du nombre d'élevages réalisés et de la quantité en kg de crevettes pêchées.

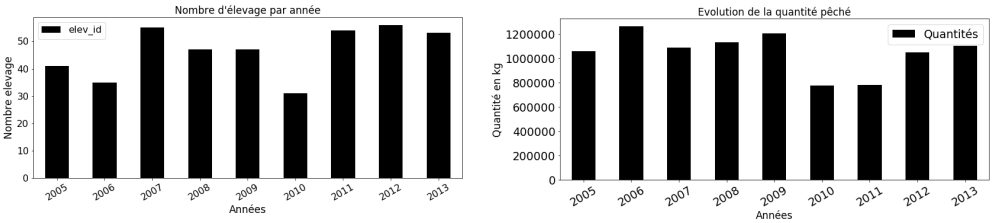

FIG. 1: Evolution du nombre d'élevage par année

Les données de qualité provenant de la SOPAC sont stockées dans un fichier au format xls qui fournit pour les productions des élevages :

— Le pourcentage de défauts visibles sur un échantillon représentatif. Il existe différents défauts relevés sur l'animal qui peuvent être situés sur la tête et le corps.

— La proportion des différents calibres existants. Il y a 8 calibres différents (16/20, 21/25, 26/30, 31/40, 41/50, 51/60, 61/80, +81). Un calibre 16/20 signifie qu'il faut 16 à 20 crevettes pour obtenir un kg de crevettes.

A noter que dans cette étude, nous utiliserons les données de qualité du premier prélèvement de chaque élevage.

2.3 Qualité des données étudiées

Dans cette étude, les différents types de données sont statiques, temporelles et acquises à différentes échelles spatio-temporelles.

Une analyse temporelle multivariée des données n'est pas directement réalisable car ces données sont parfois imprécises et manquantes et résultent d'une acquisition de données non homogène entre les fermes. Par exemple, certaines fermes privilégient certains paramètres plus ou moins régulièrement, alors que d'autres préfèrent suivre d'autres paramètres.

3 Méthodologie générale

La démarche globale de notre projet qui a pour objet d'intégrer l'ensemble des processus à l'origine des performances des élevages est présentée dans la figure 2. Seules deux étapes seront abordées dans ce papier.

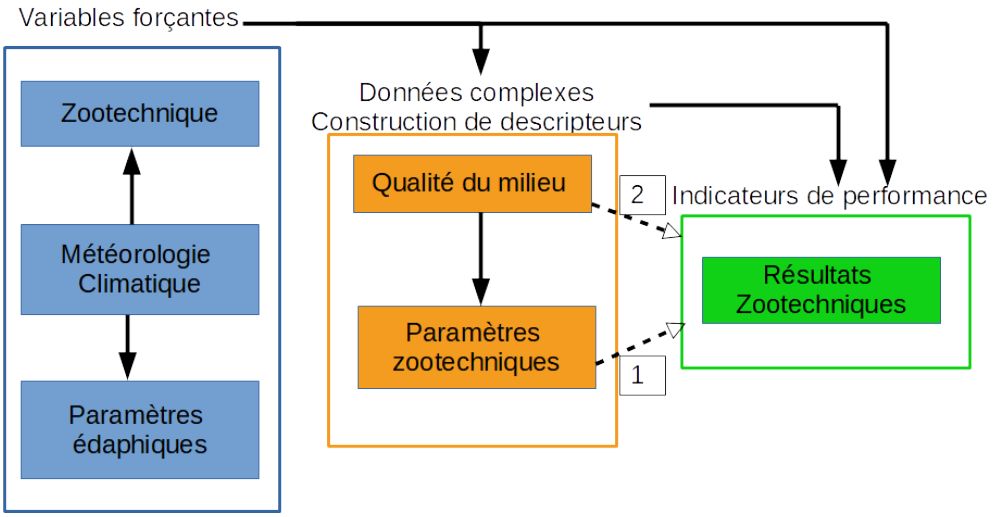

FIG. 2: Processus d'analyse mis en place pour l'étude de données issues d'une filière aquacole tropicale

L'étape 1 consiste à modéliser le lien entre l'évolution des poids mesurés au cours de l'élevage avec des indicateurs de performance (qualité finale du produit et survie). Nous prenons en compte la série temporelle des poids moyens durant l'élevage pour étudier son éventuel effet sur la performance. Cette série temporelle a été modélisée par une fonction de croissance adaptée qui permet d'extraire de nouveaux descripteurs zootechniques (évolution et vitesse de convergence de la croissance)

L'étape 2 a pour objet d'analyser la part de l'effet des conditions environnementales sur ces mêmes performances. Les données temporelles ont été analysées par la méthode *K-shape* (Paparrizos et Gravano, 2016) qui crée des clusters homogènes et bien séparés prenant en compte les formes des séries temporelles tout en les comparant.

4 Contribution

Dans cette section nous décrirons les méthodes utilisées et les résultats obtenus dans le processus de la figure 3 qui détaille les clustering réalisés dans les étapes 1 et 2.

Nous considérons ici que nous sommes dans le cadre d'une classification multi-label. En effet, la performance de chaque pêche est estimée en fonction du pourcentage répartie des défauts sur les animaux et la proportion des différents calibres. Calibres et défauts deviennent donc les labels pour la classification.

Dans l'étape 1, nous avons comparé la performance de deux techniques de classification multi-label basées sur un *classifier chain* qui prend en compte la dépendance entre les labels sélectionnés (PCC *Probabilistique classifier chain* et ECC *Ensemble classifier chain*) avec une autre technique ne prenant pas en compte la dépendance entre les labels (BR *Binary Rele-*

vance). Chaque technique est une adaptation des méthodes de classification mono-label. Pour chaque méthode multi-label, nous avons donc utilisé en entrée les classifieurs mono-label suivants : *Decision trees*, *Random Forest*, *Nearest neightbourg* et *SVM*.

Il existe dans la littérature de nombreuses approches concernant la classification multi-label (Assia, 2018). Dans (Tsoumakas et Katakis, 2009), les auteurs les regroupent en deux grandes catégories. La première catégorie englobe les méthodes qui adaptent les algorithmes mono-label 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 (Zhang et Zhou, 2014) les auteurs différencient également les méthodes selon qu'elles considèrent ou non les dépendances possibles entre les labels.

Dans notre premier modèle (étape 1), les labels sont les défauts et les calibres, discrétisés en classes de fréquences égales.

Dans l'apprentissage multi-label, les performances prédictives optimales sont obtenues par des méthodes considérant explicitement les dépendances possibles entre les labels (Dembczyński et al., 2010b). Les notions de corrélation et de dépendance entre les labels ont été discutées dans (Dembczyński et al., 2012). Par exemple, la méthode PCC *Probabilistique classifier chain* (Dembczyński et al., 2010a), proposée par le même auteur, est une technique 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 probabilité des labels intégrés.
L'avantage des méthodes de type classifieur chain CC est la vitesse d'apprentissage et la formulation des corrélations entre les labels. Cependant, la définition aléatoire de l'ordre d'apprentissage des labels reste une faiblesse (Assia, 2018). Dans le cas de notre étude, la corrélation entre les labels étudiés dépend des variables forçantes comme la température.

Dans la figure 3, la finalité des étapes est d'inclure l'ensemble des variables dans une classification multi-label.

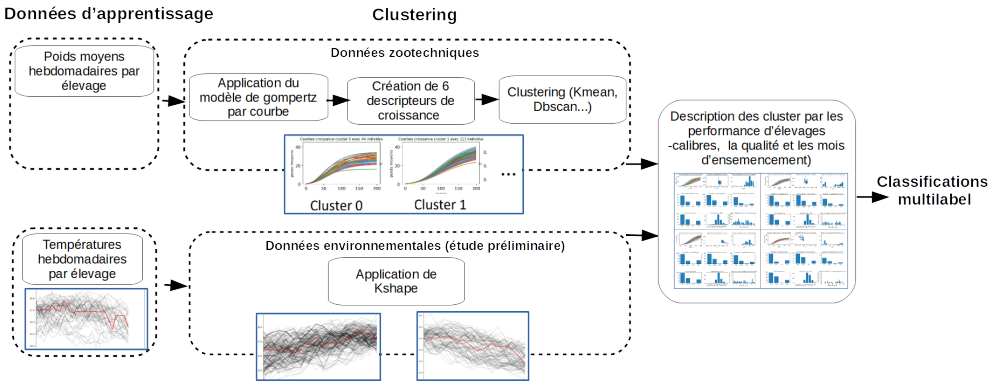

FIG. 3: Étapes du processus d'analyse des données temporelles issues de la filière crevetticole Calédonienne

4.1 Etape 1 : effet de la croissance sur la performance

4.1.1 Construction de descripteurs

Nous étudierons les courbes de croissance selon un modèle de croissance mathématique adapté. Dans la littérature, il existe plusieurs modèles mathématiques qui permettent de modéliser la croissance des animaux en aquaculture. Nous pouvons par exemple citer les modèles de Von Bertalanfly (Ferreira et al., 2007) et de Gompertz. Dans notre étude, nous avons utilisé le modèle de Gompertz car il est couramment utilisé pour modéliser la croissance des crevettes (Tjorve et Tjorve, 2017). Appliqué aux courbes de croissance des élevages étudiés, ce modèle permet de déterminer des paramètres interprétables et utilisables pour la classification supervisée ou non supervisée (clustering).

La figure 4 compare graphiquement le modèle aux données réelles et montre l'adéquation du modèle aux tendances réelles de la courbe de croissance. Pour chacune de courbe de croissance étudiée, le coefficient de détermination R^2 du modèle est corrélé à plus de 0,98 entre la courbe et son modèle. Le modèle de Gompertz est donné par la fonction de croissance

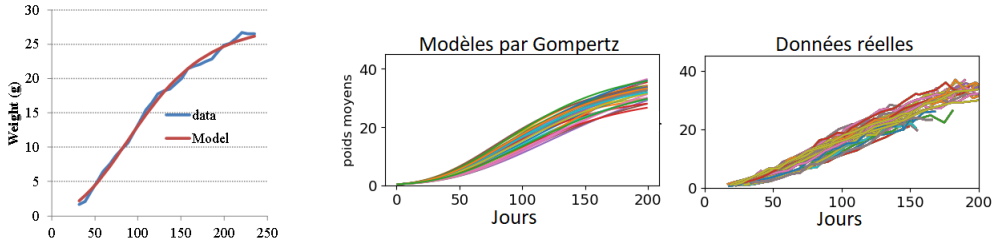

FIG. 4: Comparaison du modèle de Gompertz avec les données brutes

suivante :

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

B mesure (cf. figure 5) l'étalement du phénomène de croissance sur l'axe des abscisses et C la vitesse de convergence vers la croissance finale. Le paramètre B sera 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. Ce point varie significativement selon les valeurs conjointes des deux paramètres dans le modèle de Gompertz.

Pour obtenir les paramètres B et C, nous utilisons la bibliothèque "easynls" de R permettant de les estimer par la méthode des moindres carrées.

Afin d'augmenter le nombre de descripteurs pertinents, nous calculons également :

— Le temps correspondant au point d'inflexion P_I déterminé par la condition nécessaire $G''(t) = 0$.

— $G(t) = 1$ et $G(t) = 5$ (notés G_1 et G_5) qui fournissent la durée d'élevage pour que les animaux atteignent les poids de 1g et de 5g. Ces valeurs correspondent à des étapes clés pour les éleveurs. Le poids de 1g correspond à l'instant du début de suivi des paramètres mesurés et le poids de 5g correspond à l'instant pour lequel la vitesse de croissance devient élevée.

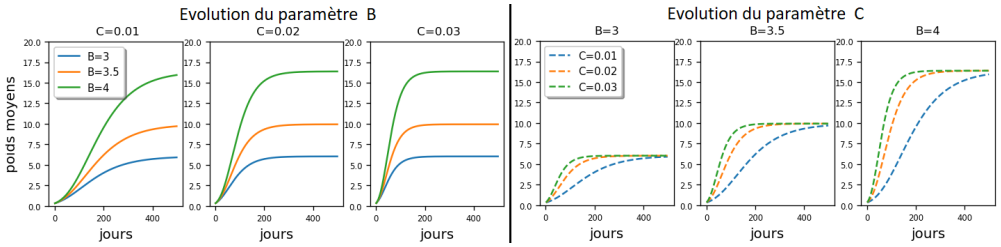

FIG. 5: Evolution de la croissance en fonction des paramètres b et c

Enfin, en plus des 5 descripteurs calculés (B, C, P_I, G_1 et G_5), nous utiliserons la durée totale d'un élevage D_e. Tous ces descripteurs seront utilisés comme attributs en entré dans l'ensemble des méthodes de classification décrites dans les sections suivantes.

4.1.2 Sélection des labels

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 :

— d_1 : "tête rouge"
— d_2 :"tête éclatée crue"
— d_3 :"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 ciblerons également la survie pour estimer les performances d'élevage.

4.1.3 Classification supervisée et non supervisée

Dans l'étape 1, nous décrivons les élevages en recherchant des groupes similaires en fonction des 6 descripteurs de croissance. Pour cela, plusieurs méthodes de clustering ont été testées (*x-means, k-means* et *dbscan*) avec plusieurs valeurs de paramètres d'entrées qui sont propres à chacune d'elles (nombre de clusters k, densité des clusters..).

Les clusters ont ensuite été décrits par des variables de qualité statiques et temporelles afin de déterminer les liens possibles entre les descripteurs de croissance et les données de performance.

4.1.4 Résultats expérimentaux

La figure 6 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 croissance de leurs individus (i.e élevage). Ces typologies

caractérisent 5 pratiques d'élevages de la filière, qui peuvent être décrites par des variables forçantes (mois d'ensemencement, âge du bassin...).

Les méthodes *x-means* et *dbscan* ont été utilisées et comparées, mais ne seront pas présentées car elles fournissent des résultats comparables avec la méthode *k-means*.

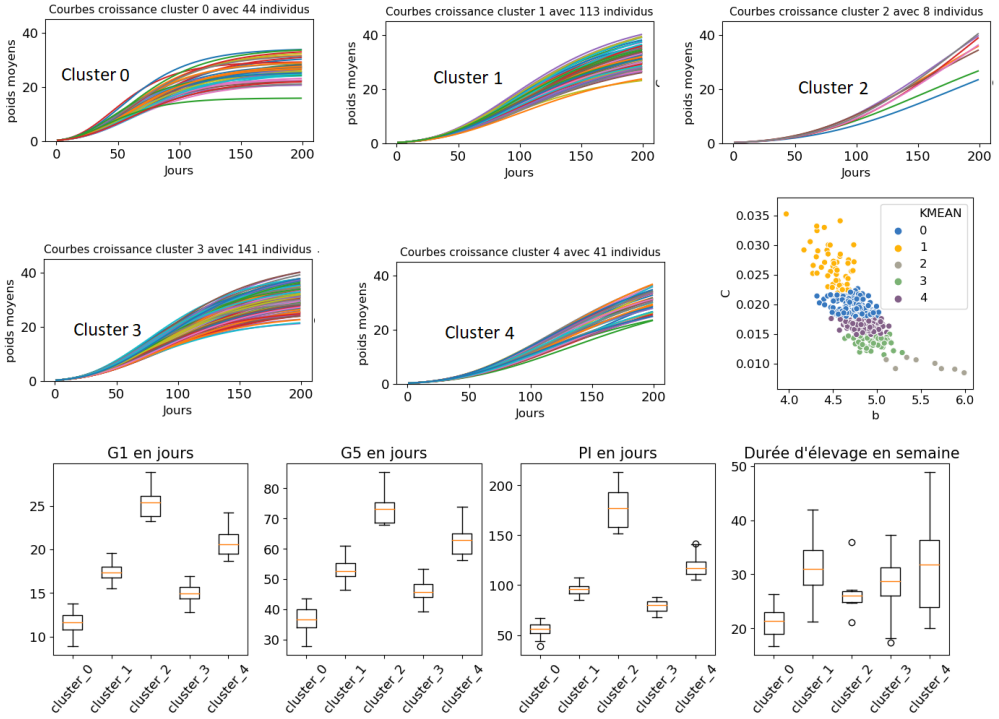

FIG. 6: Principaux groupes obtenus par les descripteurs de croissance

La figure 7 décrit les clusters selon les périodes d'ensemencement des élevages. D'après cette figure, les courbes de croissance qui convergent rapidement vers leur valeur finale sont associées à des élevages ensemencés en début d'année (cluster 0). La Nouvelle-Calédonie étant située dans l'hémisphère Sud, ces élevage débutent donc pendant la période chaude.

Ainsi, la différence de distribution des valeurs des descripteurs entre le cluster 0 et les clusters 1 et 4 peut être expliquée par une température saisonnière plus élevée pour le cluster 0 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 correspond donc à une survie les plus faibles et un taux de têtes éclatées le plus élevé et donc des performances d'élevage très faibles.

La figure 8 montre la répartition moyenne des défauts et de la survie dans chacun des clusters. La figure 9 présente la répartition des différents calibres des élevages par cluster.

Hormis le cluster 2, dont les individus sont clairement marqués par une croissance particulièrement lente, nous pouvons remarquer que la distribution des calibres dans les clusters

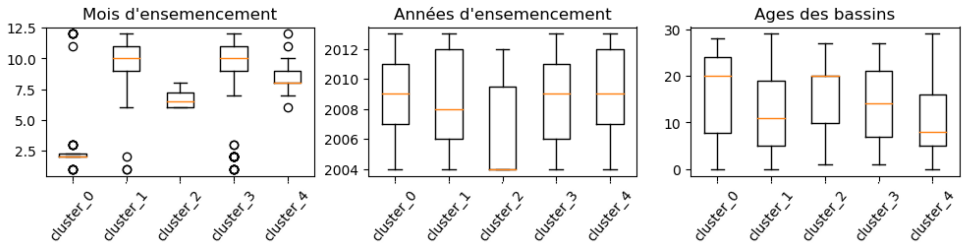

FIG. 7: Description des clusters par des variables temporelles

suivent une distribution gaussienne autour du calibre 41/50. Par exemple, le cluster 0 a en moyenne 10% d'individus de calibre 31/40, plus de 50% d'individus de calibre 41/50, et 25% d'individus du calibre 51/60.

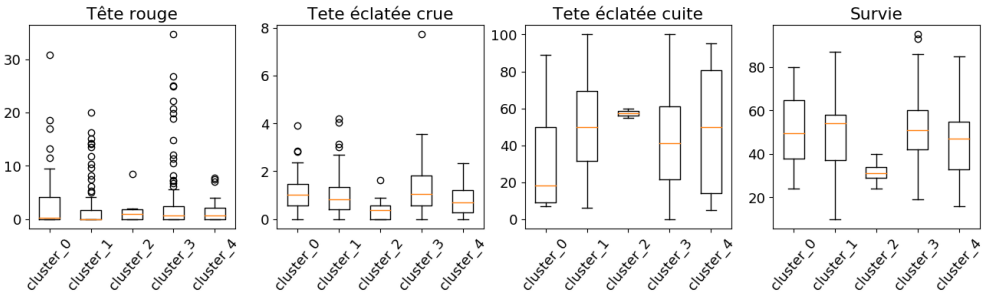

FIG. 8: Qualité des groupes d'élevages

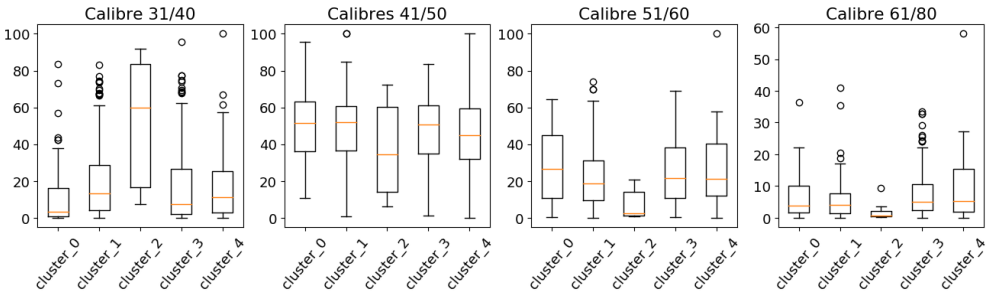

FIG. 9: Calibres des groupes d'élevages

Nous avons pu décrire statistiquement chaque cluster par les paramètres de performance de manière indépendante qui a permis d'expliquer certaines pratiques mais il est très difficile de

Classifieur Multi-label	Classifieur de base	Recall	Precision	F1-score
ProbabilisticClassifierChain	RDMForest	0,988091902	0,969453641	0,978634538
EnsembleClassifierChain	RDMForest	0,986013823	0,947366918	0,966241634
BinaryRelevance	RDMForest	0,966435037	0,957196423	0,961718328
ProbabilisticClassifierChain	KNN	0,779107828	0,775219572	0,777108913
EnsembleClassifierChain	KNN	0,779107828	0,775219572	0,777108913
BinaryRelevance	KNN	0,781023408	0,774466472	0,777606493
ProbabilisticClassifierChain	DecTree	0,741773385	0,757221842	0,748859903
EnsembleClassifierChain	DecTree	0,733741256	0,75909815	0,745339159
BinaryRelevance	DecTree	0,733741256	0,75909815	0,745339159

TAB. 1: Comparaison des performances des classifieurs multi-label sur des labels dépendants

donner une description conjointe de ces paramètres pour pouvoir les prédire. Pour cela, nous avons considéré les mêmes données avec les mêmes attributs et plusieurs labels par individus que nous avons construit à partir des données de performance disponibles dans la base de données de la SOPAC.

Le tableau 1 présente les performances de plusieurs classifieurs multi-label avec les 6 descripteurs construits ci-dessus et en ciblant les différents labels sélectionnés (calibres, défauts et survie). Globalement, les méthodes considérant les dépendances entre les labels (PCC, ECC) ont une meilleure performance que la méthode BR. Nous constatons que la méthode ECC *Ensemble classifier chain* qui construit plusieurs classificateurs en chaîne avec un ordre d'étiquettes aléatoire possède la performance la plus élevée. Ainsi, l'ordre d'arrivée des cibles dans le classifieur pourrait impacter la performance du modèle.

4.2 Étape 2 : analyse de la tendance de la température

Les différents clusters obtenus dans la section précédente ont pu montrer une relation entre croissance et mois d'ensemencement ce qui s'explique par une influence de la température d'eau des bassins sur la croissance des animaux comme l'ont montré (Jackson et Wang, 1998).

Dans le cadre de notre étude visant à intégrer la qualité du milieu dans le modèle (cf figure 2), nous avons effectué une première analyse des évolutions temporelles de température de l'eau des bassins, prises sur la totalité de la période des élevages, par la méthode de clustering dite *k-shape*.

Les clusters obtenus par la méthode *k-shape* ont été décrits en fonction de la survie des crevettes par élevage. Tout comme la méthode *k-mean*, *k-shape* impose de fournir un nombre (k) de clusters. 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. La méthode de *k-shape* fournit des clusters de mêmes formes mais ne prend pas en compte leurs écarts d'amplitudes.

Nous avons donc, pour chaque cluster, montré en figure 10, affiché en rouge les séries de températures les plus élevées i.e. au dessus de la température médiane dans le cluster (2ème quartile) et en bleu les températures les plus fraîches i.e. en dessous de la série médiane. Pour les courbes rouges des deux clusters, la survie est en moyenne supérieure à 55% et est meilleure

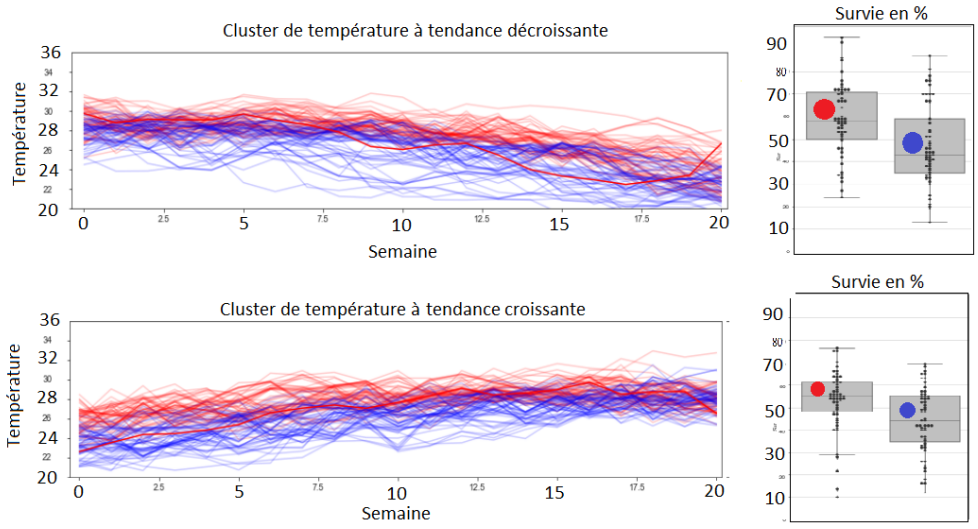

FIG. 10: Clustering des séries temporelles de température par la méthode *k-shape*

que la survie des courbes bleues. Cependant, nous pouvons remarqué que le taux de survie est plus élevé dans le cluster de température décroissante présentant un 3ème quartile à 70% contre 60% pour le second cluster. Une température élevée en début d'élevage permettrait donc d'obtenir des élevages avec une survie plus élevée.

5 Conclusion et perspective

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 papier, nous avons contribué à deux étapes de cette méthodologie.

L'étape 1 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 décrits 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-label sur les mêmes données avec les mêmes descripteurs mais considérant des labels de plusieurs données de qualité 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, les classifieurs multi-label qui considèrent les relations possibles entre les labels ont eu de meilleures performances (precision, recall...) et notamment *Ensemble classifier chain* qui construit plusieurs classificateurs en chaîne avec un ordre d'étiquettes aléatoire. La définition aléatoire de l'ordre d'apprentissage des labels reste la faiblesse de ces classifieurs. En perspective, nous proposons d'étudier de plus près ce problème.

L'étape 2 du processus, qui consiste en une première étude, a permis de mettre en évidence le lien entre l'évolution de la température durant la totalité de l'élevage avec la survie.

Pour compléter le processus d'analyse montré en figure 2, nous proposons en perspective d'analyser globalement 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...).

Remerciements

Nous remercions le Groupement des Fermes Aquacoles (GFA) et la Société des Producteurs Aquacoles Calédoniens (SOPAC) de la Nouvelle-Calédonie de nous avoir fourni les données.

Références

Assia, B. (2017-2018). *Contribution en apprentissage semi-supervise sous contexte multi-label*. Ph. D. thesis, Oran.

Bourke, G., F. Stagnitti, et B. Mitchell (1993). A decision support system for aquaculture research and management. *Aquacultural Engineering 12*(2), 111 – 123.

Charles, D. (1979). Presentation coordonnée de differents modeles de croissance. *Revue de Statistique Appliquee*, 5–22.

Cheng, Z., F. Flouvat, et N. Selmaoui-Folcher (2017). Mining recurrent patterns in a dynamic attributed graph. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 631–643. Springer.

Czogaa, E. et T. Rawlik (1989). Modelling of a fuzzy controller with application to the control of biological processes. *Fuzzy Sets and Systems 31*(1), 13 – 22.

Dembczyński, K., W. Cheng, et E. Hullermeier (2010a). Bayes optimal multilabel classification via probabilistic classifier chains. pp. 279–286.

Dembczyński, K., W. Waegeman, W. Cheng, et E. Hüllermeier (2010b). On label dependence in multi-label classification. *Workshop proceedings of learning from multi-label data*, 5–12.

Dembczyński, K., W. Waegeman, W. Cheng, et E. Hüllermeier (2012). On label dependence and loss minimization in multi-label classification. *Machine Learning 88*(1), 5–45.

FAO (2016). Food and agriculture organization of the united nations, the state of world fisheries and aquaculture.

FAO (2018). Food and agriculture organization of the united nations. aquaculture.

Ferreira, J., L. Falconer, J. Kittiwanich, L. Ross, C. Saurel, K. Wellman, C. Zhu, et P. Suvanachai (2015). Analysis of production and environmental effects of nile tilapia and white shrimp culture in thailand. *Aquaculture 447*, 23 – 36.

Ferreira, J., A. Hawkins, et S. Bricker (2007). Management of productivity, environmental effects and profitability of shellfish aquaculture the farm aquaculture resource management (farm) model. *Aquaculture 264*(1), 160 – 174.

Giraudel, J. L., D. Aurelle, P. Berrebi, et S. Lek (2000). *Application of the Self-Organizing Mapping and Fuzzy Clustering to Microsatellite Data : How to Detect Genetic Structure in Brown Trout (Salmo trutta) Populations*, pp. 187–202. Berlin, Heidelberg : Springer Berlin Heidelberg.

Guinand, B., K. T. Scribner, A. Topchy, K. S. Page, W. Punch, et M. K. Burnham-Curtis (2004). *Sampling issues affecting accuracy of likelihood-based classification using genetical data*, pp. 245–259. Dordrecht : Springer Netherlands.

Gusmawati, N., B. Soulard, N. Selmaoui-Folcher, C. Proisy, A. Mustafa, R. L. Gendre, T. Laugier, et H. Lemonnier (2018). Surveying shrimp aquaculture pond activity using multitemporal vhsr satellite images - case study from the perancak estuary, bali, indonesia. *Marine Pollution Bulletin 131*, 49 – 60. Special Issue : Indonesia seas management.

Jackson, C. et Y.-G. Wang (1998). 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.

Jardim, G. et R. Ricardo, Luis (2016). Aquaculture production optimization through enhanced data analytics. Sem PDF 6th Offshore Mariculture Conference.

Jesus, E., A. Artifice, J. Sarraipa, G. Mcmanus, et F. Luis-Ferreira (2018). A training programme to support aquasmart project exploitation.

Joao et Rihtar (2016). Data analytics in aquaculture. *SIKDD 2016*.

Joao, R., K. Sarraipa, et V. Seferis (2016). Data analytics : models and algorithms for intelligent data analysis.

Lee, P. G. (2000). Process control and artificial intelligence software for aquaculture. *Aquacultural Engineering 23*(1), 13 – 36.

Paparrizos, J. et L. Gravano (2016). k-shape : Efficient and accurate clustering of time series. *ACM SIGMOD Record 45*, 69–76.

Rahman, A. et M. Shahriar (2013). Algae growth prediction through identification of influential environmental variables : A machine learning approach. *International Journal of Computational Intelligence and Applications 12*.

Soulard, B., J. Frappier, J. Herlin, et B. Benoit (2009). Stylog : base de données pour le suivi des élevages de crevettes de nouvelle-calédonie.

Tian, X., P. Leung, et E. Hochman (1993). Shrimp growth functions and their economic implications. *Aquacultural Engineering 12*(2), 81 – 96.

Tjorve, K. et E. Tjorve (2017). 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.

Tsoumakas, G. et I. Katakis (2009). Multi-label classification : An overview. *International Journal of Data Warehousing and Mining 3*, 1–13.

Yu, R., P. Leung, et P. Bienfang (2006). Predicting shrimp growth : Artificial neural network versus nonlinear regression models. *Aquacultural Engineering 34*(1), 26 – 32.

Zhang, M.-L. et Z.-H. Zhou (2014). A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on 26*, 1819–1837.

Zhi, C. (2018). *Mining recurrent patterns in a dynamic attributed graph. Application to aquaculture Pond Monitoring by satellite images*. Ph. D. thesis, University of New Caledonia.

Summary

The success of an aquaculture farm depends on many factors that can be zootechnic or economic. Actors involved in this complex production process must identify the best conditions to optimize the quality of the product. For example, the survival and the growth of the animals are important elements that depend on many variables that are monitoring. Thus, this field generates a lot of data that are generally under-exploited because of their complexity (heterogeneous, temporal, spatial, etc.) and come from different sources. (producer, provendiers...). In this article, we will describe the approach applied to analyze a dataset generated by multiple actors which are established on tropical farms of the *Litopenaeus stylirostris* shrimp produced in New Caledonia between 2003 and 2015. The purpose of our approach is to cross the production and marketing data to identify the best zootechnical practices and the best possible environmental conditions to optimize the efficiency of the farms. For that, we propose different methodological scenarios in data science (clustering and classification) on the data in order to (i) identify trends or groups of trends of the most optimal farm practices and to (ii) predict them. Given the complexity of the data, these models will be applied to constructed parameters. For example, from an animal growth model based on weight data measured by the farmers. We will present the results interpreted by experts and the discussion of the study.