GAST

GESTION ET ANALYSE DE DONNÉES SPATIALES ET TEMPORELLES

UNIVERSITÉ de la NOUVELLE-CALÉDONIE

INSTITUT DE SCIENCES EXACTES ET APPLIQUÉES

Ifremer

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

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

*ISEA -Institut des Sciences Exactes et Appliquées - Université de la Nouvelle calédonie

**IFREMER-LEAD Nouméa, Nouvelle-Calédonie

GOUVERNEMENT DE LA NOUVELLE-CALÉDONIE

SOPAC

G.F.A.

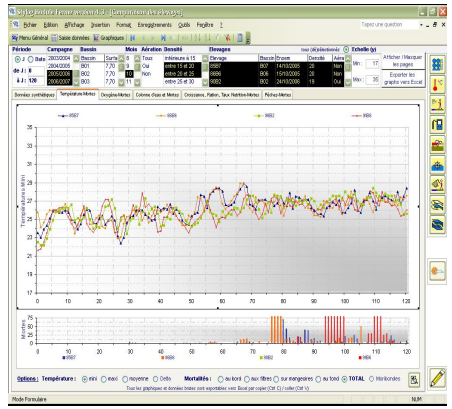
Contexte

De plus en plus de données générées par les procédés industriels agricoles.
-> un grand nombre de défis en matière d'analyse de données

L'aquaculture : Industrie majeure pour les produits de la mer (FAO, 2016, 2018)

Exemple : production aquacole Source FAO 2012 (données 2010)

- 5.7 millions de tonnes de crustacés
  - 26.8 milliards de dollars



- L'un des enjeux majeurs de la filière est l'amélioration des performances des élevages
- Quantité de données massives, complexes et spatio-temporelles -> Science des données

Informations recherchées par l'expert

Exploitation des données pour :

- Une meilleure évaluation des techniques de production
- Une meilleure adaptation au marché
- Une meilleure compréhension des facteurs responsables du développement des maladies => limiter le risque.

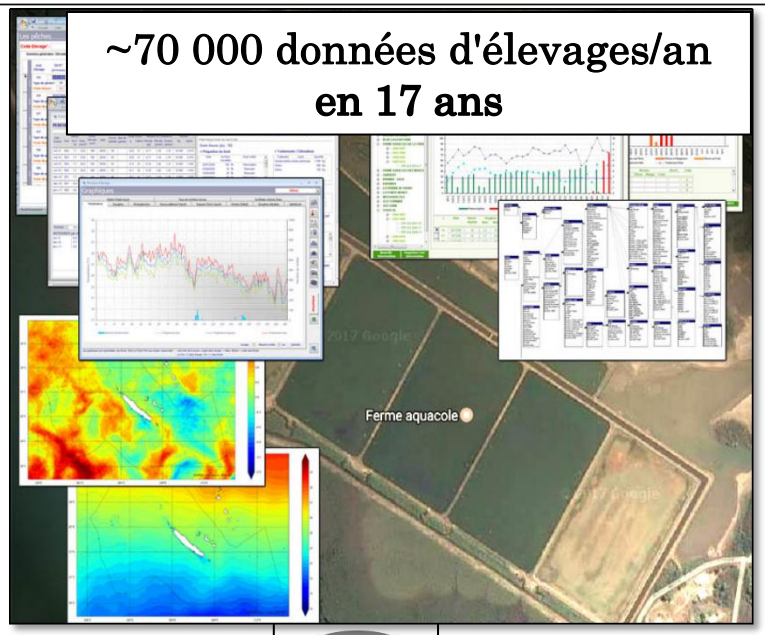
Exemple de la crevetticulture Calédonienne

+ de 600 élevages de 5 et 7 mois

- **Données d'élevage :**
  - Séries temporelles
  - Environnement zootechnie

  (ex. turbidité, salinité, température, distribution alimentation...)

~70 000 données d'élevages/an en 17 ans



- **Données de qualité du produit**
  - Défauts (branchies, tête...)
  - Différents calibres
  - Premium/Bas de gamme

+ de 5000 pêches

Exemple de la crevetticulture Calédonienne

Par élevage : 5 à 7 mois

Préparation du bassin

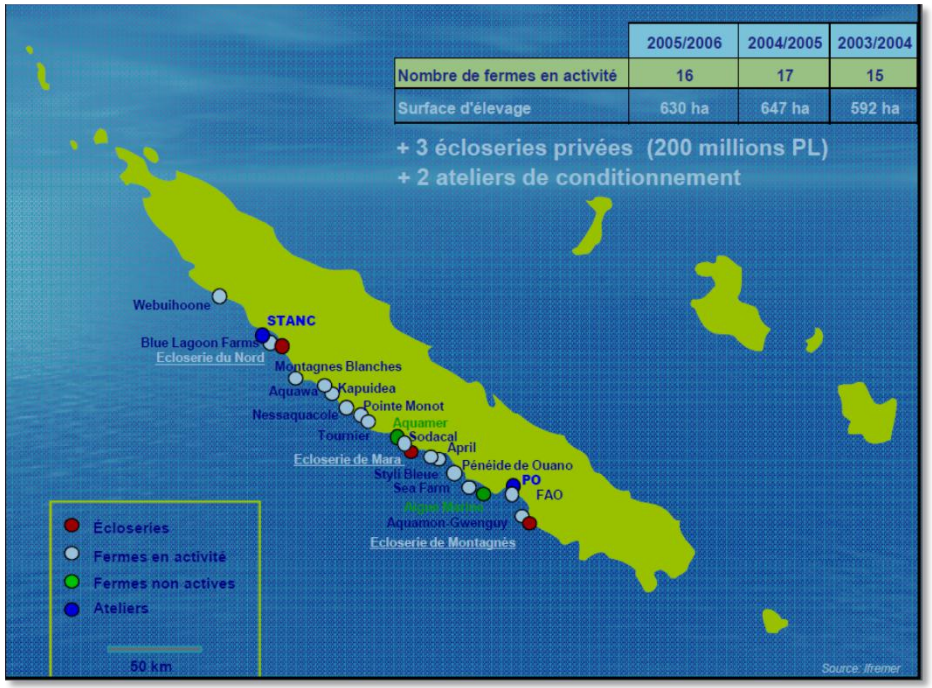
Mise en eau Ensemencement

L'élevage jusqu'à la première pêche

de la première pêche à la vidange finale

	2005/2006	2004/2005	2003/2004
Nombre de fermes en activité	16	17	15
Surface d'élevage	630 ha	647 ha	592 ha

+ 3 écloseries privées (200 millions PL)
+ 2 ateliers de conditionnement



- **Données d'élevage :**
  - Séries temporelles
  - Environnement zootechnie

  (ex. turbidité, salinité, température, distribution alimentation…)

Exemple de la crevetticulture Calédonienne

Les données de qualité, relevées par la SOPAC

- Estimation en laboratoire du pourcentage d'apparition des défauts sur un échantillon

Cassée

Tête rouge

Tête éclatée cuite

Calibre 51/60

Patte rouge

Produits décotés -> Bas de gamme

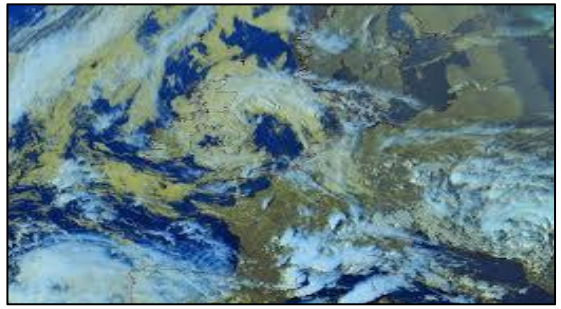
Sans défaut

Premium

Plusieurs variables ciblées

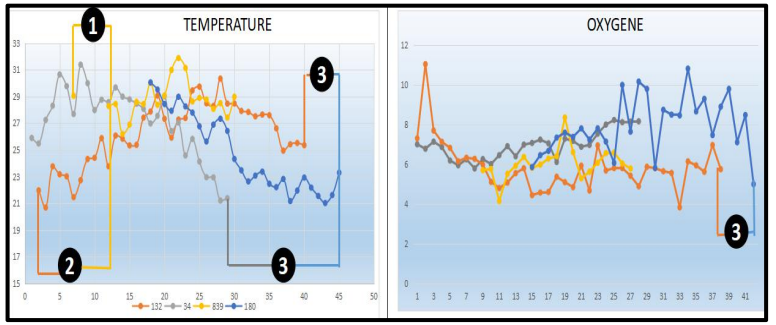
Description des données

- Variables forçantes
  - Météorologie
  - Age du bassin
  - …



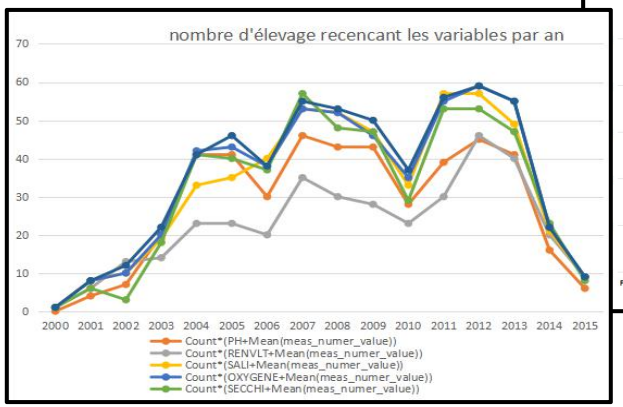
En vert : données temporelles
En bleu : données statiques

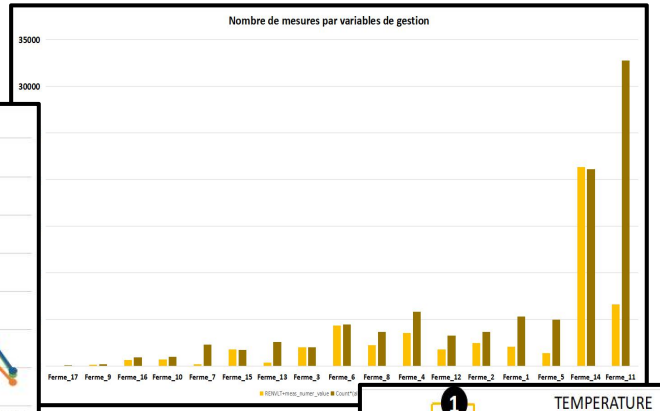
- Variables de gestion
  - Renouvellement de l'eau
  - Turbidité
  - Oxygène dissous
  - …

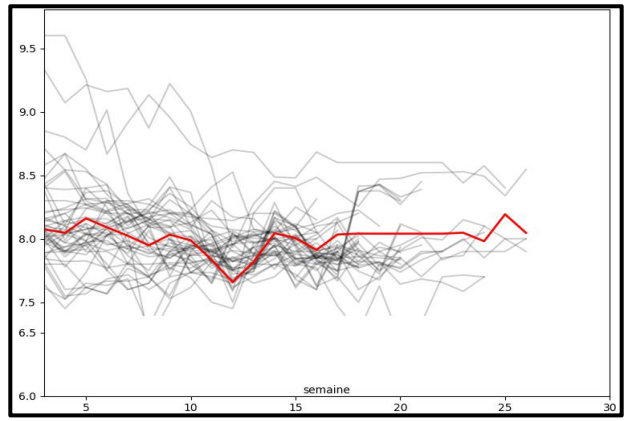


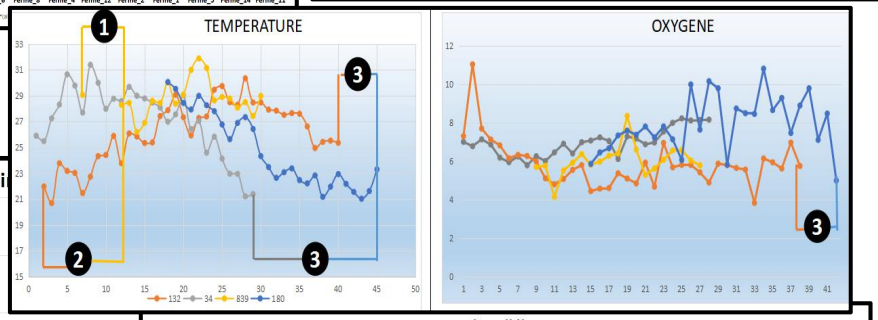
- Indicateurs de performance :
  - Ratio de conversion alimentaire
  - Taux de survie
  - Gain de poids corporel
  - Rendement
  - …

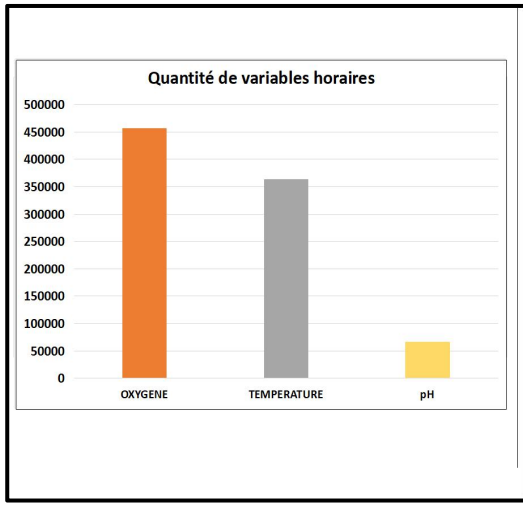
Complexité des données

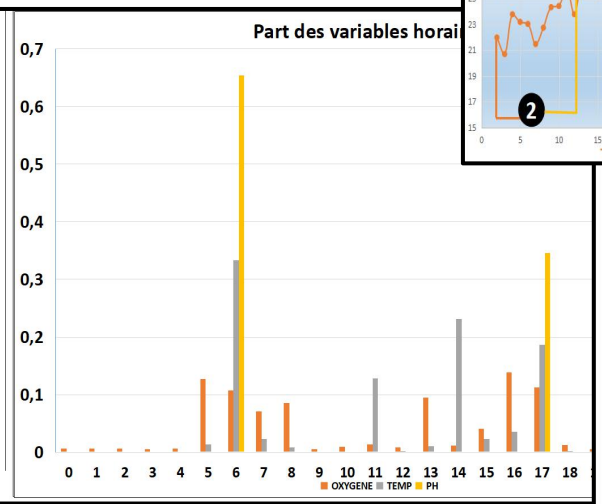


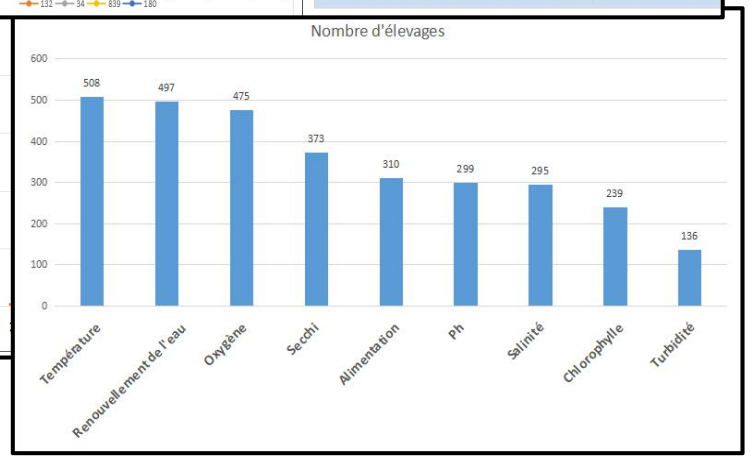












Objectifs

Développer une démarche en science des données pour répondre aux questionnements de l'expert

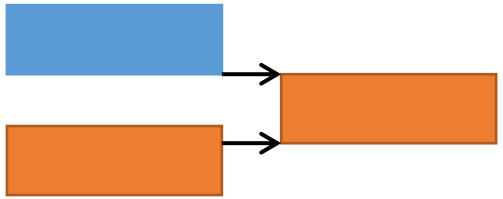
- S'affranchir de la complexité des données (hétérogènes, spatio-temporelles, multivariées,...)
- Croiser les données de production avec les données de qualité

- Créer des outils d'aide à la décision

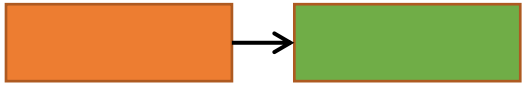
Etat de l'art

Variables forçantes | **Variables de gestion** | **Performances**

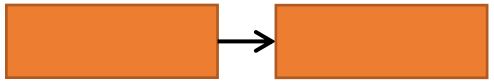
Jesus, E., A. Artifice, J. Sarraipa, G. Mcmanus, et F. Luis-Ferreira (2018). *A training programme to support aquasmart project exploitation*.



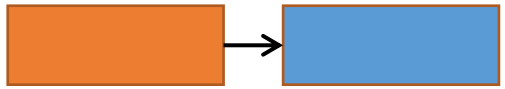
Bourke, G., F. Stagnitti, et B. Mitchell (1993). *A decision support system for aquaculture research and management*. Aquacultural Engineering 12(2), 111 – 123. (feature selection)



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.

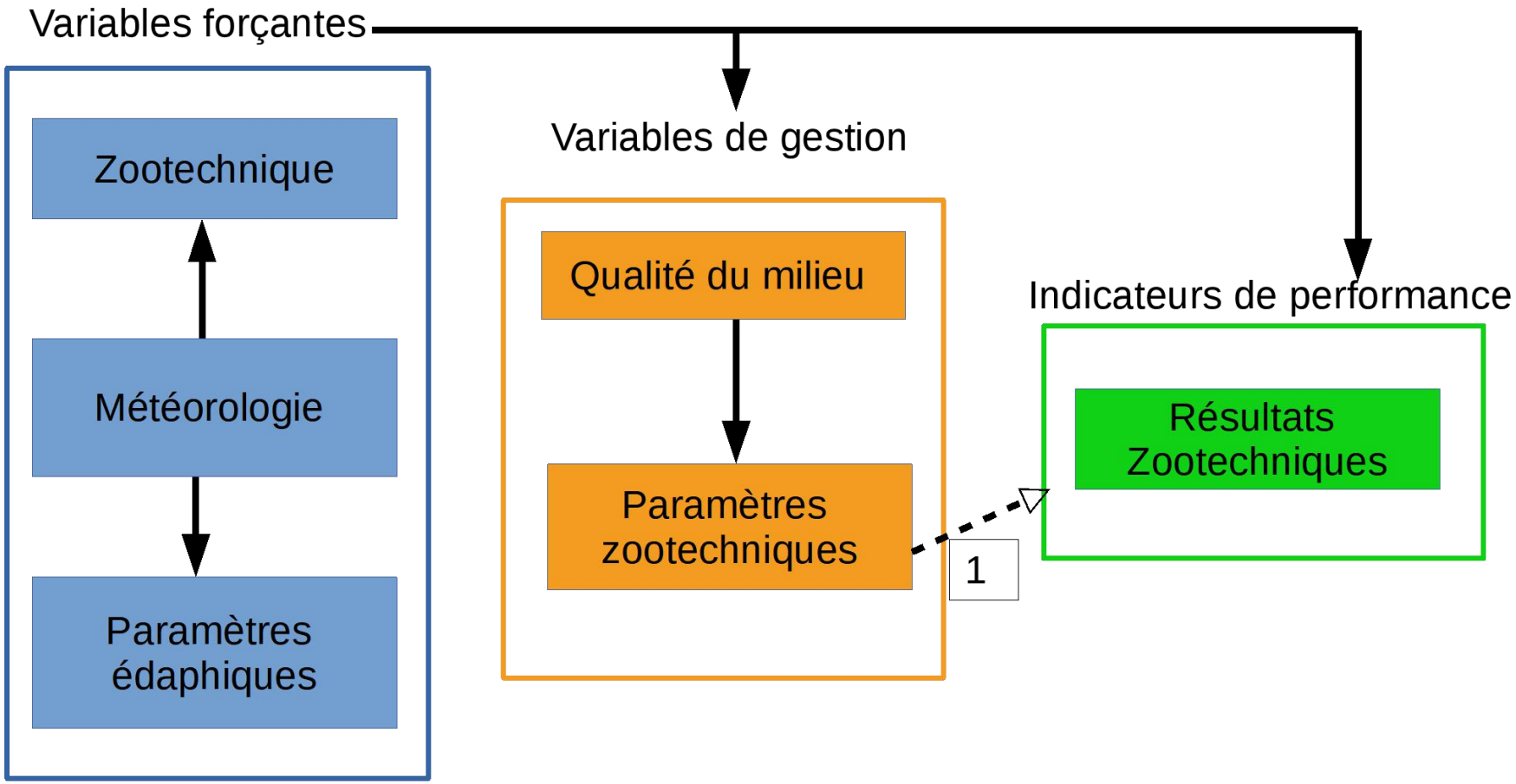


Ferreira, J., L. Falconer, J. Kittiwanich, L. Ross, C. Saurel, K. Wellman, C. Zhu, et P. Suvanachai (2015). *Analysis of production and environmental effects of nile tilapia and white shrimp culture in thailand*. Aquaculture 447, 23 – 36.



Contribution méthodologique

Une méthodologie réalisable à partir des données d'étude



1 : *Une approche mise en place sur une échelle temporelle jamais atteinte*

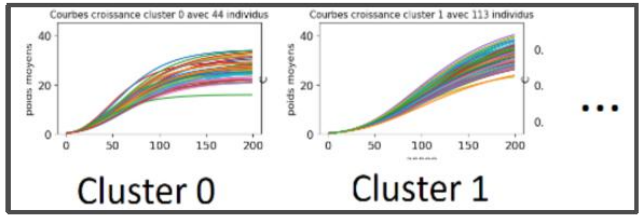
Description de l'étape 1

Données d'entrée

Poids moyens hebdomadaires par élevage

Création de descripteurs de croissance

Application du modèle de gompertz par courbe

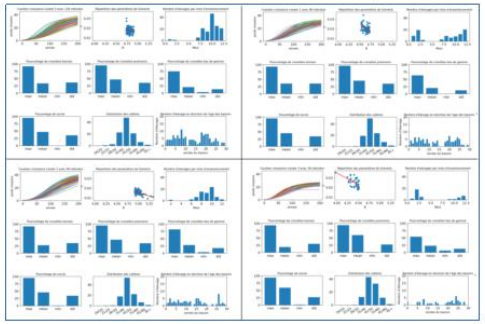


Clustering

Comparaison de plusieurs méthodes (Kmean, Dbscan...)

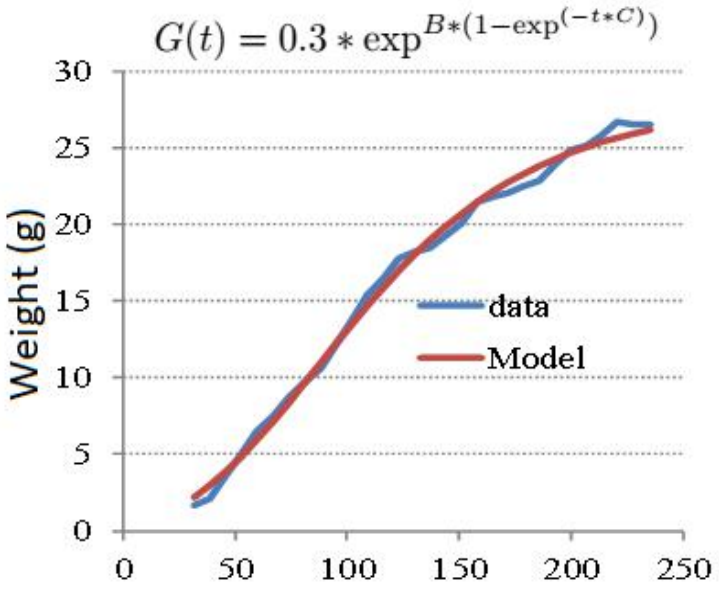
Analyse des clusters

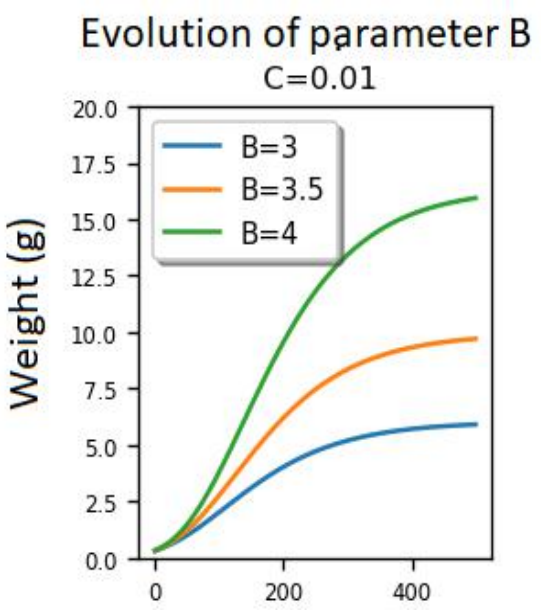
Description par les performances d'élevages (calibres, la qualité et les mois d'ensemencement)

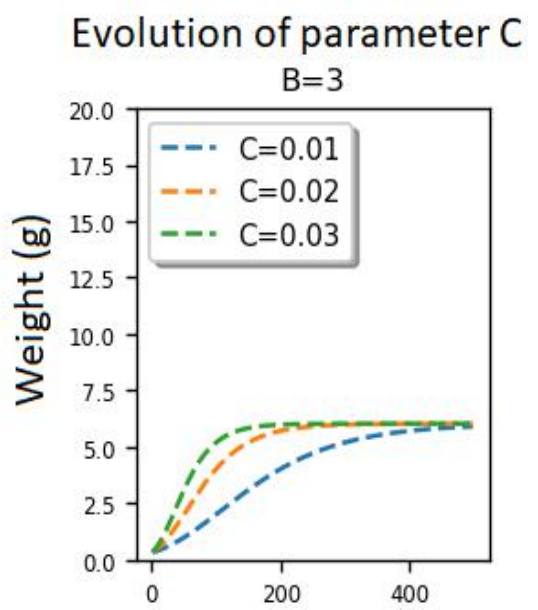


Génération de 6 nouveaux descripteurs zootechniques

Modèle de Gompertz tester sur chaque élevage avec $R^2 > 0.98$







- B : Mesure l'étalement du phénomène de croissance sur l'axe des abscisses
- C : la vitesse de convergence vers la croissance finale
- *PI* : Point d'inflexion (G''(x) = 0)
- *G1* : la durée d'élevage pour atteindre le poids d'1g (G(x) = 1)
- *G5* : la durée d'élevage pour atteindre le poids de 5g (G(x) = 1)
- *De :* la durée d'un élevage

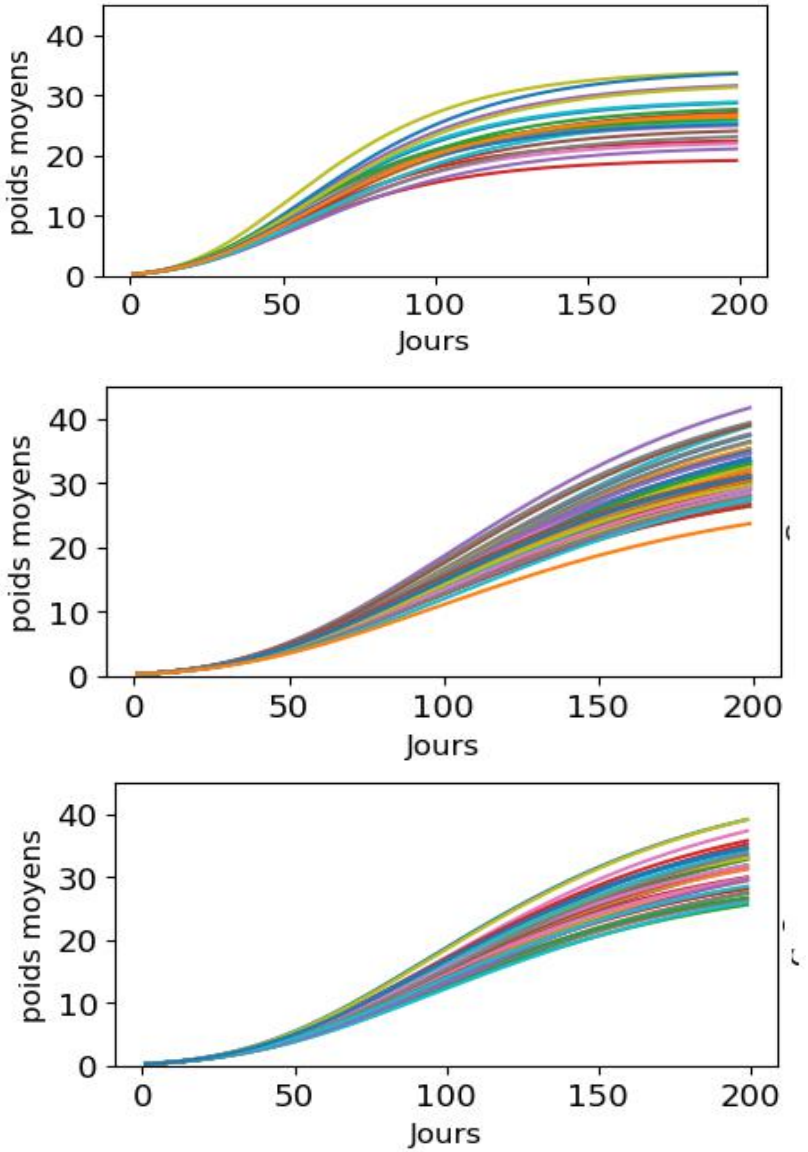
Clustering des descripteurs de croissance

- Utilisation de plusieurs méthodes de clustering
  - K-mean
  - DBScan
  - X-mean

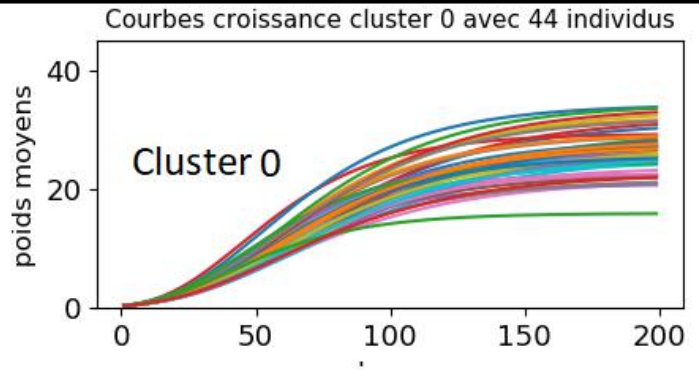
Les différentes méthodes fournissent des clusters comparables.

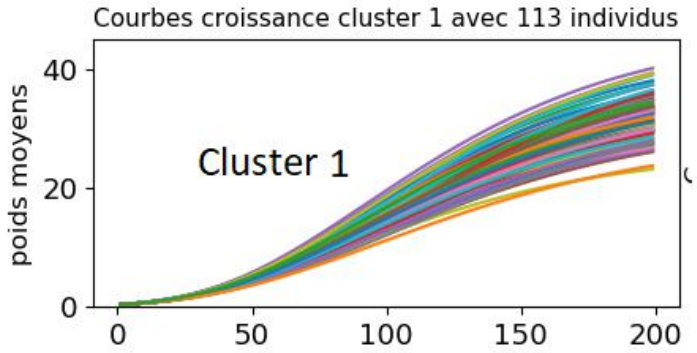
Seule la méthode K-mean est présentée ensuite

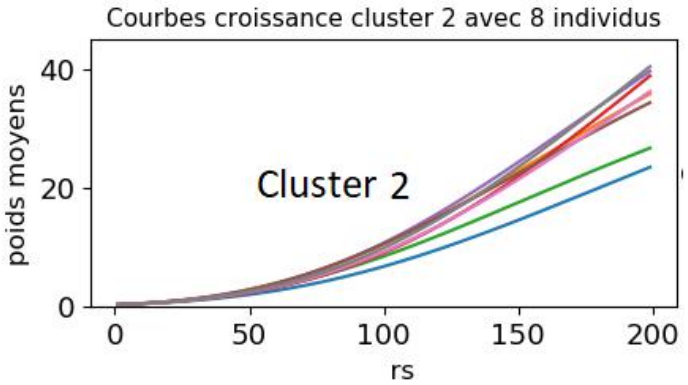
Exemple de clusters obtenus par les 3 méthodes

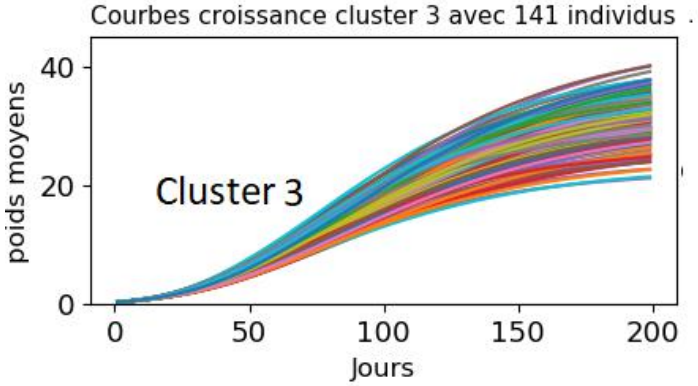


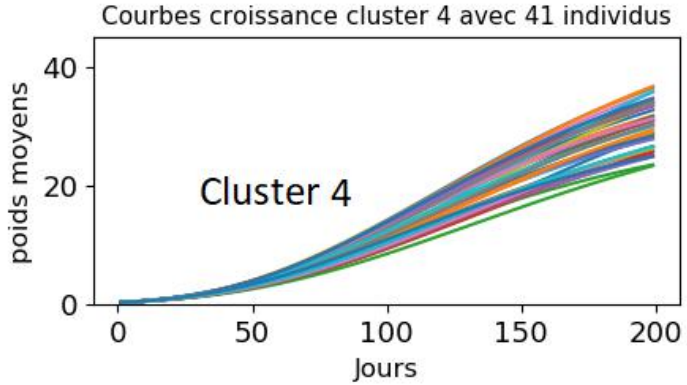
Résultats du Clustering







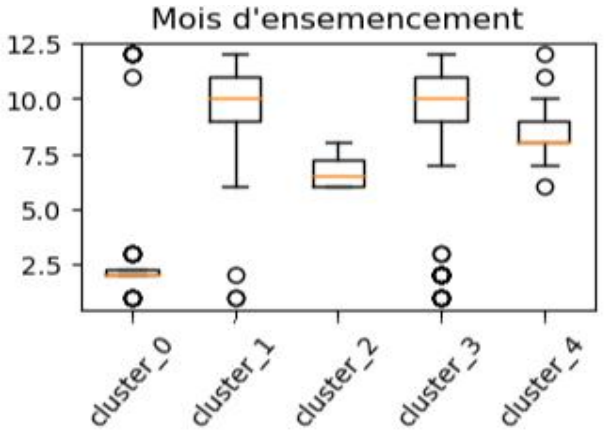


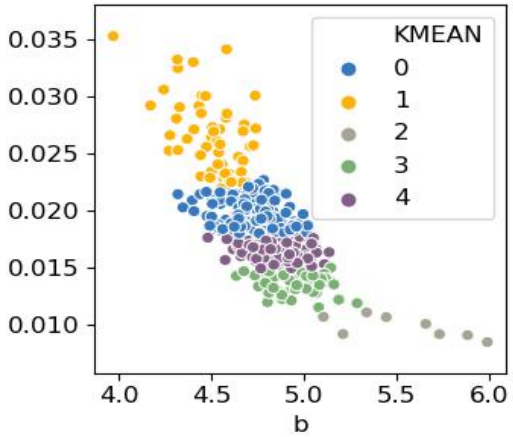


Rappel :

B : Mesure l'étalement du phénomène de croissance sur l'axe des abscisses

C : la vitesse de convergence vers la croissance finale



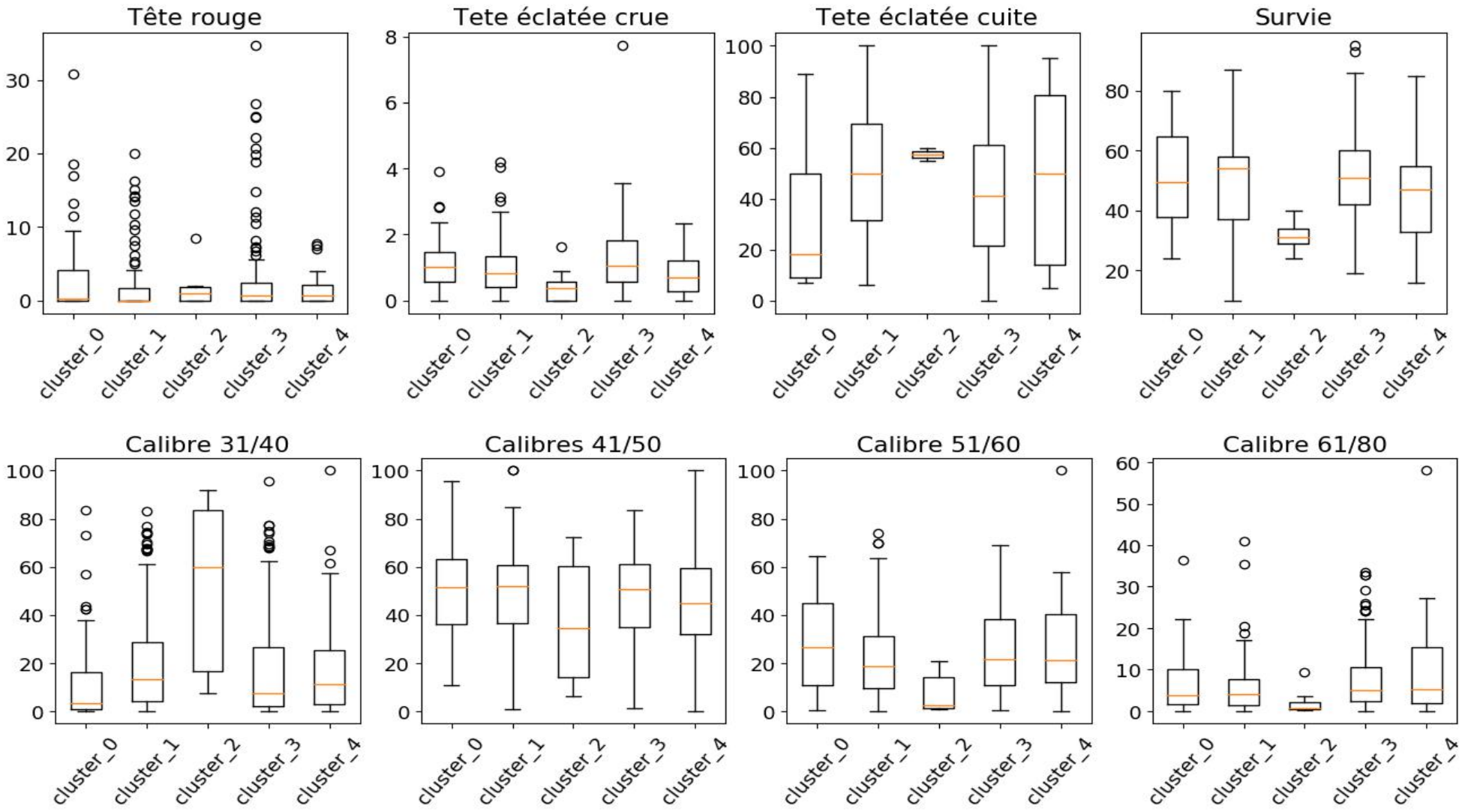


L'analyse des variables de production montre qu'en fonction du mois d'ensemencement les paramètres varient fortement

- Mise en place de normes de croissance à l'échelle de la filière selon le mois d'ensemencement

Description des clusters par les données de qualité

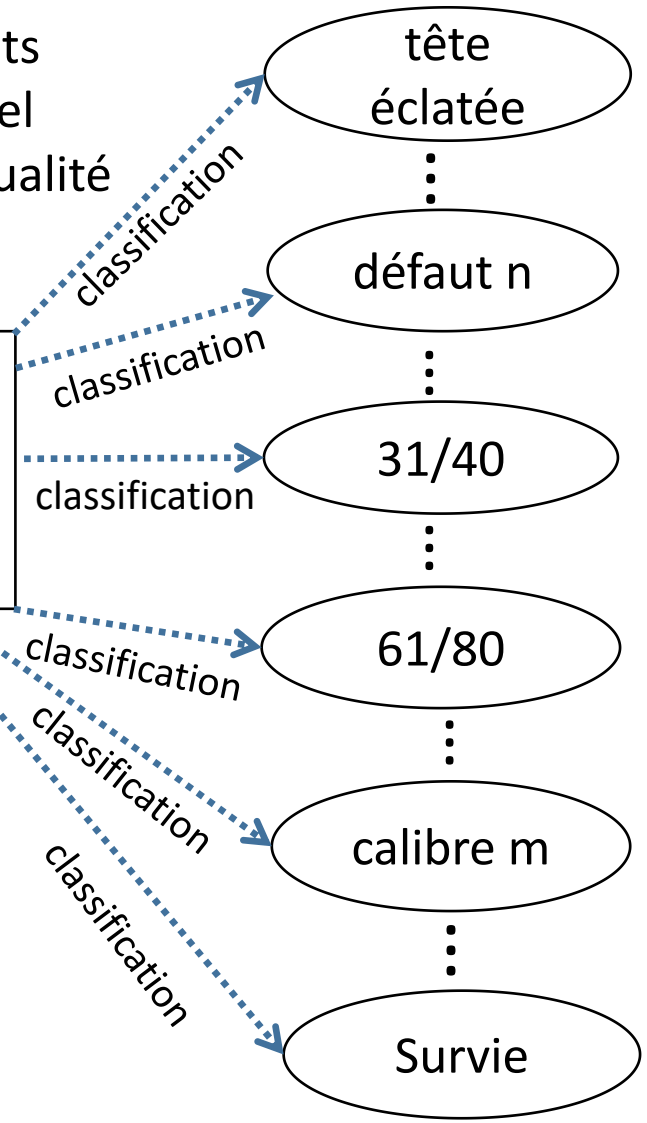
Données de qualité



Classification supervisée

Utilisation de différents classifieurs mono-label sur chaque variable de qualité

- MLP neural_network
- KNeighbors
- Decision Tree
- Random Forest



Performances classifieurs

	F1-score	Recall
tête éclatée	0.4725	0.4371
défaut n	...	...
31/40	0.5525	0.4569
61/80	0.5850	0.6352
calibre m	...	...
Survie	0.4950	0.4412

Approche monolabel inadaptée

- Classification multi-label

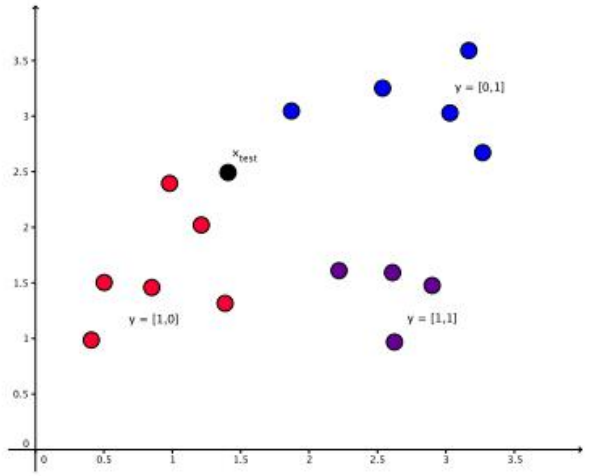
Classification multi-label

Deux approches concernant la classification multi-label

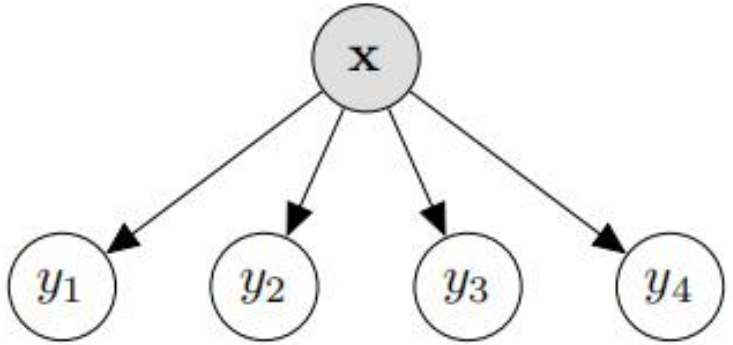
Tsoumakas, G. et I. Katakis (2009). *Multi-label classification : An overview. International Journal of Data Warehousing and Mining* 3, 1–13.

- Les méthodes qui adaptent les algorithmes mono-label pour traiter directement des données multi-label.

$$\hat{y}_j = \begin{cases} 1 & \left(\frac{1}{k}\sum_{i|\mathbf{x}^{(i)}\in \mathrm{Ne}(\tilde{\mathbf{x}})} y_j^{(i)} > 0.5\right) \\ 0 & \text{otherwise} \end{cases}$$



- Les méthodes qui transforment un problème mono-label en un problème multi-label.



Classification multi-label

Deux techniques de classification sont comparées

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.

Considèrent la dépendance entre les labels

- PCC Probabilistique Classifier Chain
- ECC Ensemble Classifier Chain

X	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0

X	Y_1	Y_2
$\mathbf{x}^{(1)}$	0	1
$\mathbf{x}^{(2)}$	1	0
$\mathbf{x}^{(3)}$	0	1
$\mathbf{x}^{(4)}$	1	0
$\mathbf{x}^{(5)}$	0	0

X	Y_1	Y_2	Y_3
$\mathbf{x}^{(1)}$	0	1	1
$\mathbf{x}^{(2)}$	1	0	0
$\mathbf{x}^{(3)}$	0	1	0
$\mathbf{x}^{(4)}$	1	0	0
$\mathbf{x}^{(5)}$	0	0	0

X	Y_1	Y_3	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1

La définition aléatoire de l'ordre d'apprentissage des labels reste une faiblesse.

Ne considère pas la dépendance entre les labels

- BR Binary Relevance

X	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0

X	Y_2
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

X	Y_3
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

X	Y_4
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	1

Classification multi-label

Les 6 descripteurs zootechniques

ID_elevage	B	C	PI	G1	G5	De
1	4.5	0.025	50	15	35	20
2	5.5	0.25	100	20	50	30
	...	...	...	...	...	...
400	5	0.20	125	25	70	25

Défauts et les calibres discrétisés en classes de fréquence égale.

Survie	Calibre 16/20	Calibre 21/30	....	Tête éclatée cuite
0	0	1	...	0
1	0	1	...	0
...	...	...	...	...
1	1	1	...	0

Résultats des classifications multi-label

Multi-label Classifier	Base Classifier	Recall(%)	Precision(%)	F1-score(%)
PCC	RDMForest	98.8	96.9	97.9
ECC	RDMForest	98.6	94.7	96.6
BR	RDMForest	96.6	95.7	96.2
PCC	KNN	77.9	77.5	77.7
ECC	KNN	77.9	77.5	77.7
BR	KNN	78.1	77.4	77.8
PCC	DecTree	74.2	75.7	74.9
ECC	DecTree	73.4	75.9	74.5
BR	DecTree	73.4	75.9	74.5

- Dépendance des labels : méthodes en chaine plus performantes
  - Déterminer le meilleur ordonnancement des labels pour optimiser les performances des classifications
  - Pour l'expert : comprendre plus finement l'interdépendance des labels

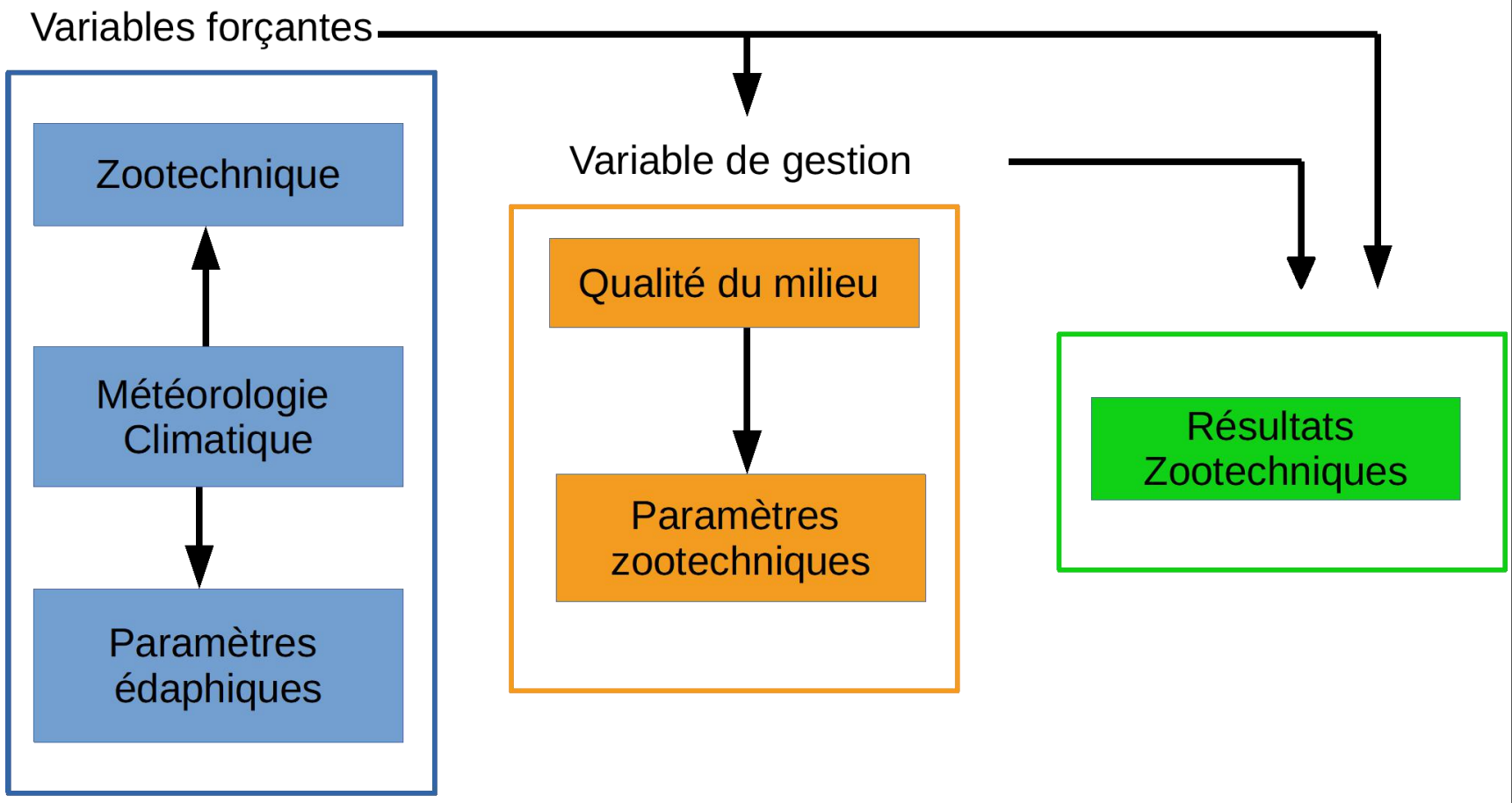
Conclusion

- Analyse des séries de données spatio-temporelles relevées habituellement dans les procédés industriels à destination de la production (qualité du milieu, taux de croissance...), dans le but d'en améliorer les performances.
  - *Des données complexes, imprécises, parfois incomplètes*
  - Filière crevetticole Calédonienne : croisement de 2 bases de données (GFA : données de production et SOPAC : données de qualité)
    - *Données zootechniques, environnementales, performances d'élevage pour le suivi et la gestion de la filière (600 élevages)*

- Un processus est proposé pour intégrer l'ensemble des données acquises
  - Une première étude : 6 descripteurs générés grâce au modèle de croissance de Gompertz .
    - Clustering des descripteurs et analyses des clusters
      - ✓ possibilité de créer des normes de croissance de la filière selon le mois d'ensemencement mais aussi de qualité (création d'un indice de performance)
      - ✓ possibilité d'un futur couplage avec les données économiques
    - Modèle supervisé par plusieurs données de qualité, certaines sont dépendantes entre elles ou dépendent d'autres variables temporelles (saison, climat...), et impactent les performances d'élevage.
    - Meilleure performance des classifieurs multilabel qui prennent en compte les dépendances entre les labels

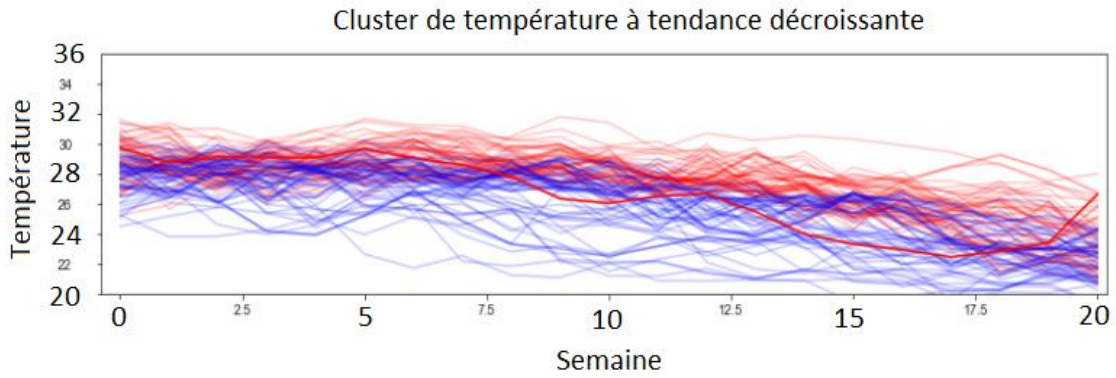
Perspectives

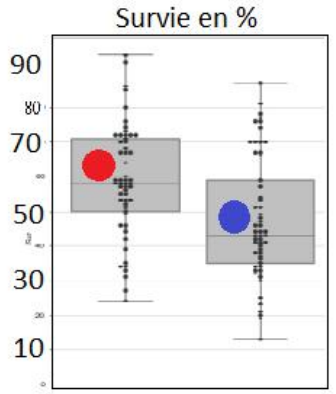
- Intégrer les variables forçantes et de qualité du milieu dans une classification supervisée par les résultats de performance

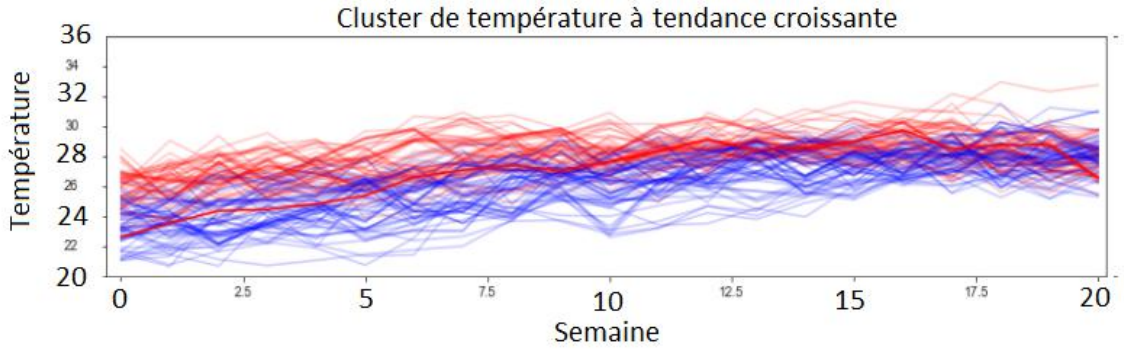


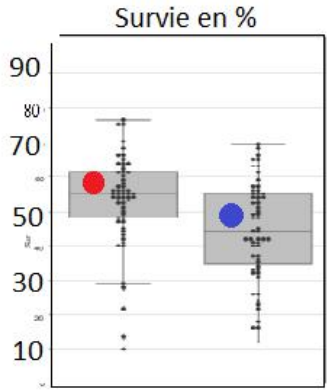
Perspectives

- Clustering des données environnementales par la méthode K-shape
  - Ex : Cluster de la température prise sur la durée totale de l'élevage









- Clustering des séries temporelles sur différentes fenêtres de temps
  - Ex : considérer la température avant le Point d'Inflexion