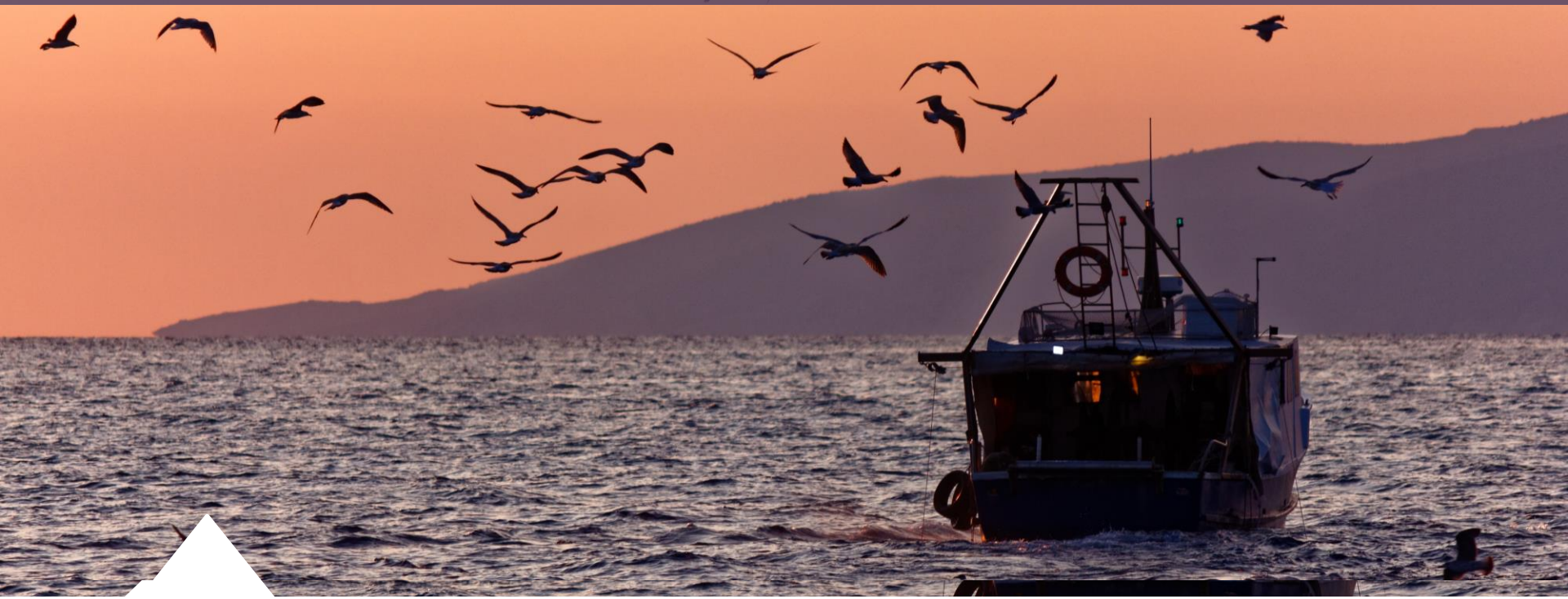
Stratégies de pêche

Tâm Le Minh et Sylvain Moinard

22/01



université PARIS-SACLAY

AgroParisTech

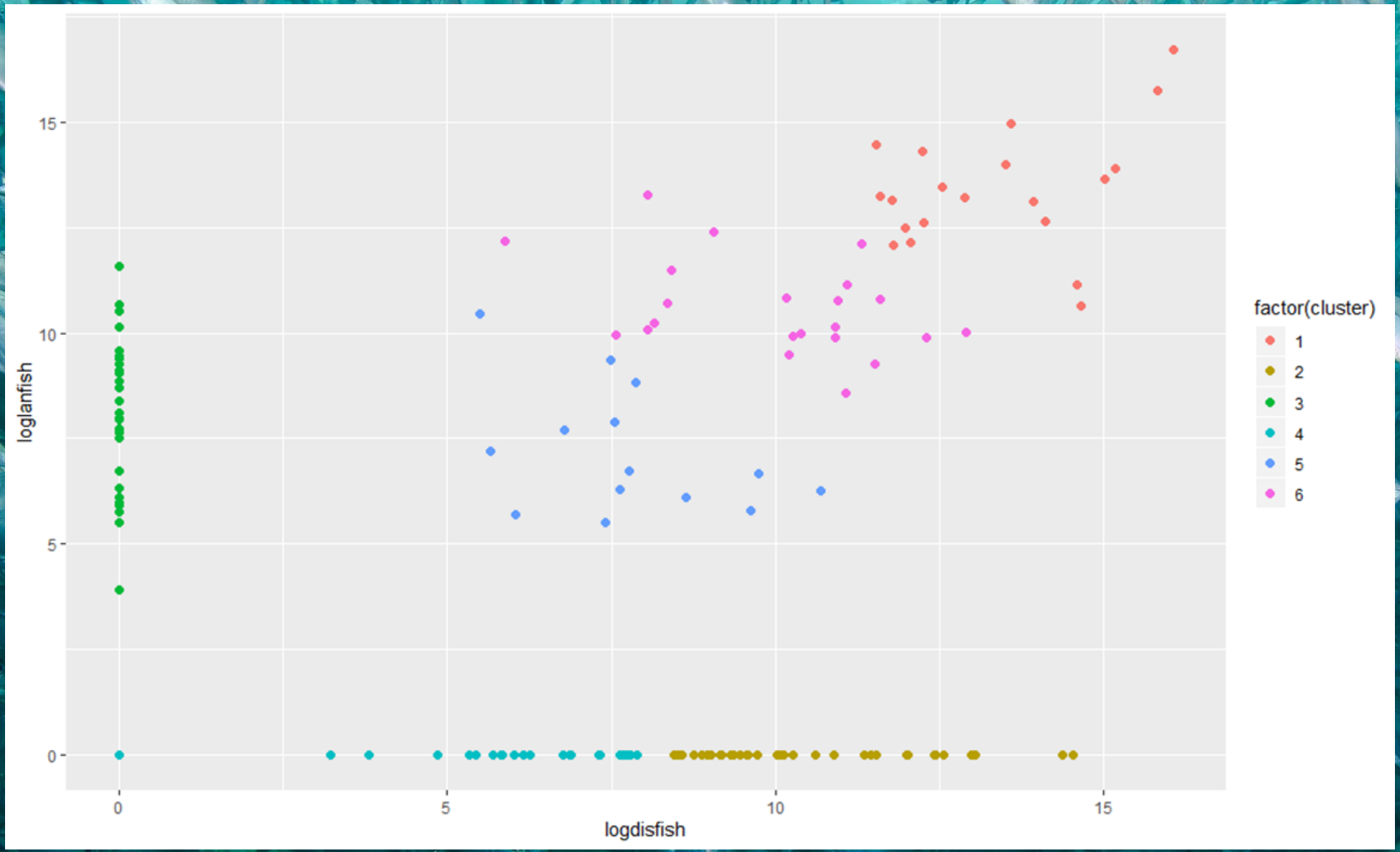
Ifremer

Contexte
Interdiction des rejets en mer

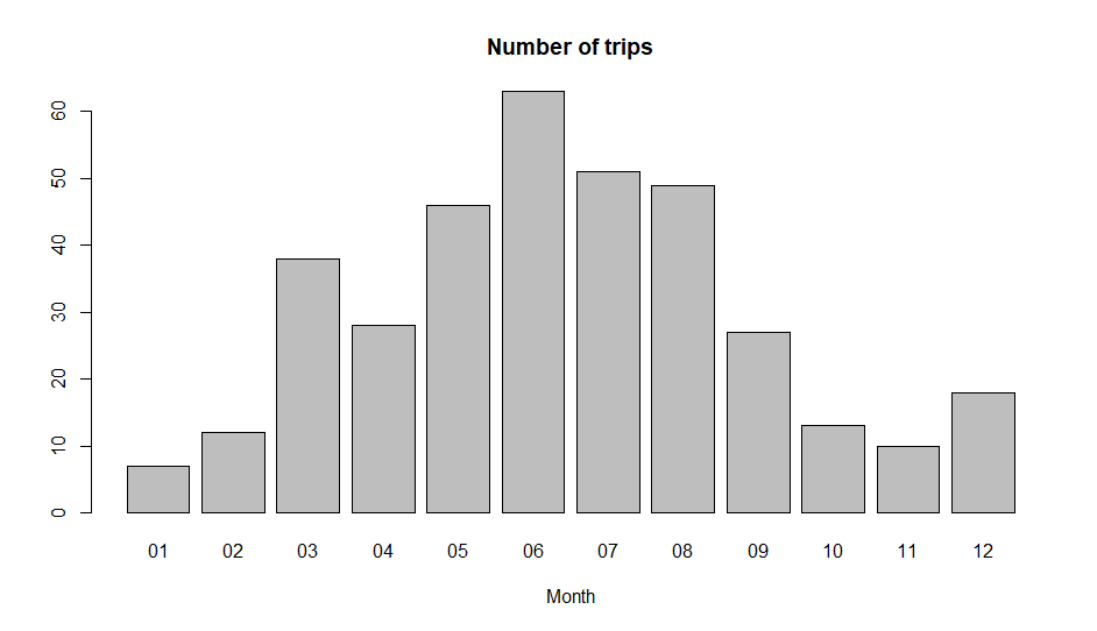
Données pour **150 espèces** :
quantités rejetées et débarquées

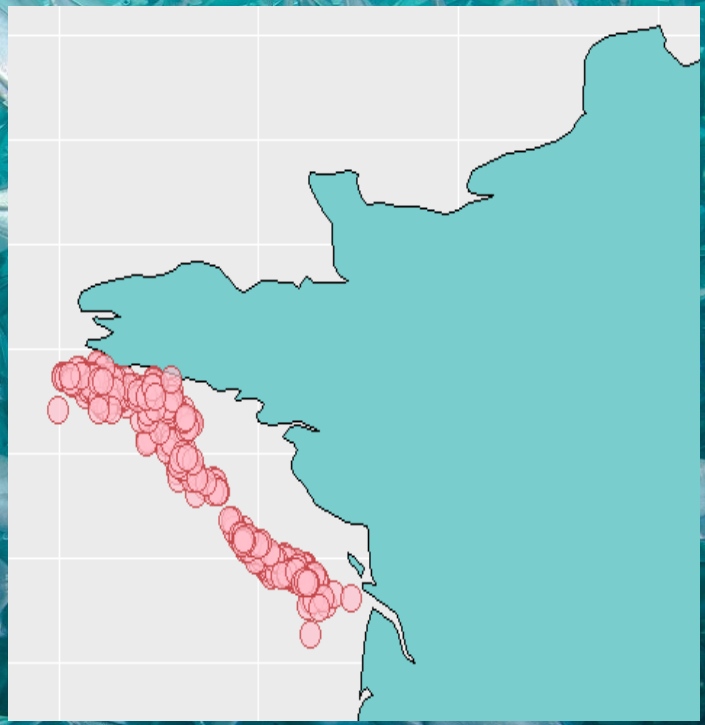
Objectif : prédire les **quantités rejetées**
Signal multivarié

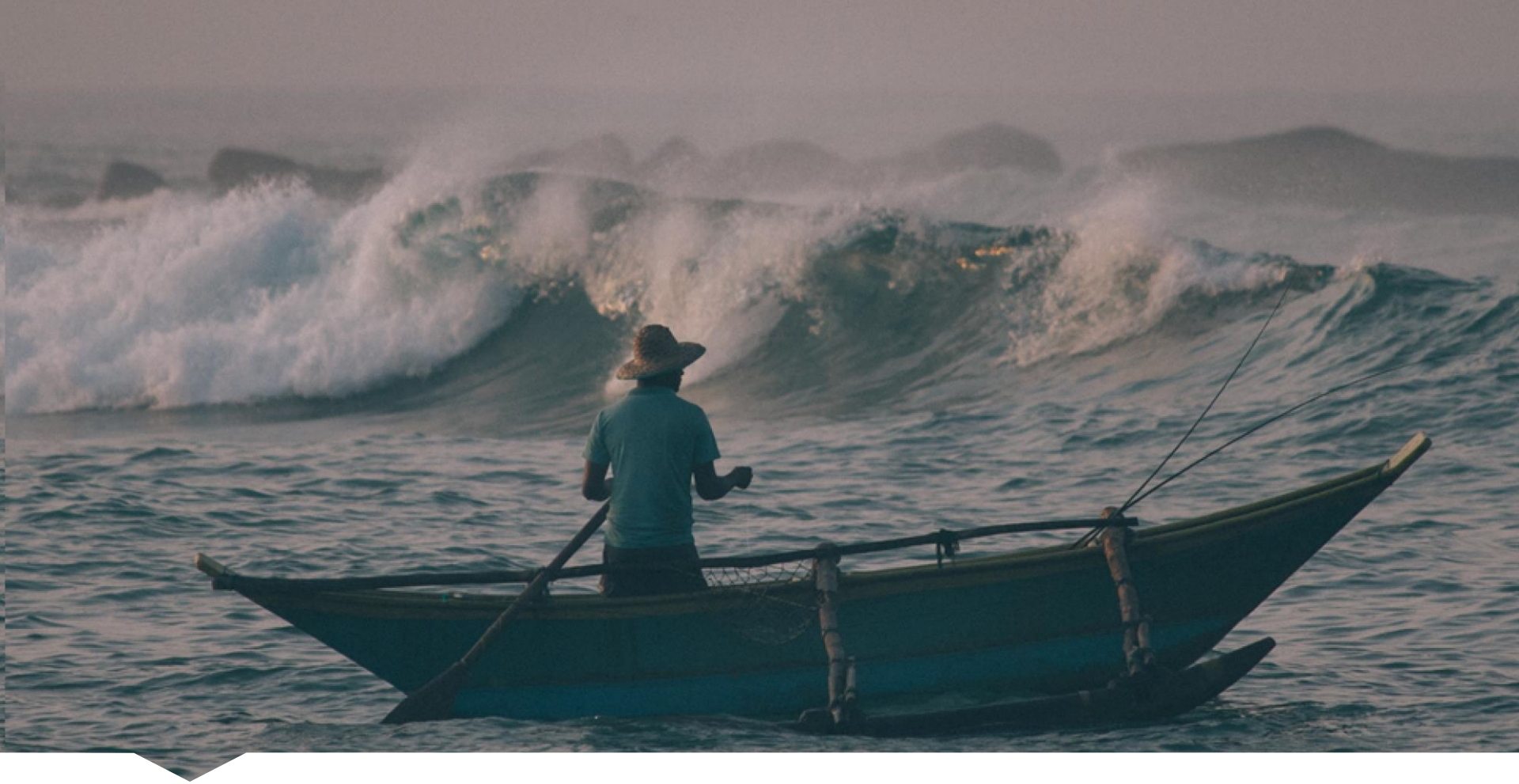
Données
Aperçu des quantités pêchées



Données
Répartition spatio-temporelle







Approche naïve

Régressions

n = 362, p = 150, corrélation entre les variables

Régression linéaire

- Système couplé

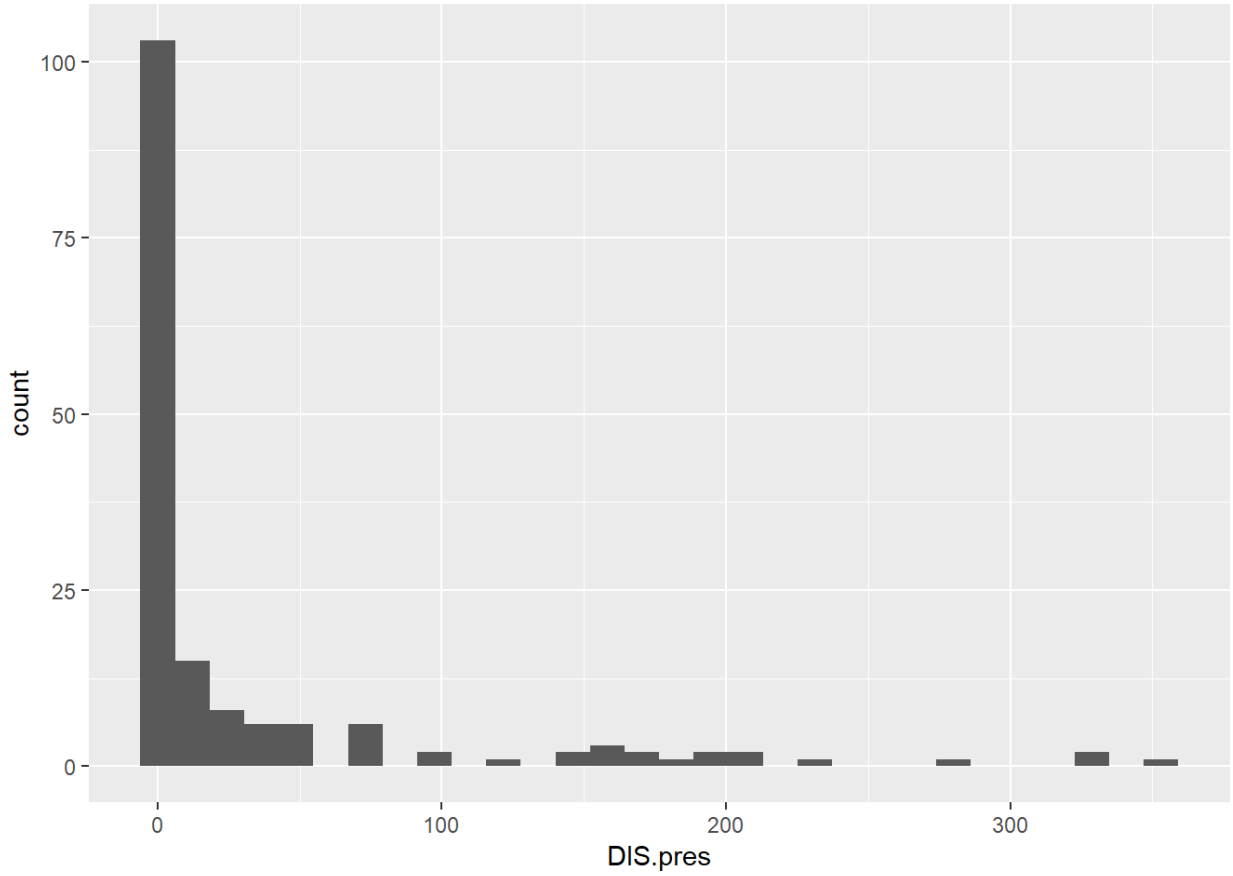
$$Y_i = \sum_{j \neq i} \alpha_j^i Y_j + \sum_j \beta_j^i X_j + Cofacteurs + Intercept$$

$$D.(Y_1, Y_2, \ldots, Y_n)^t = \beta.(X_1, X_2, \ldots, X_n)^t + Cofacteurs + Intercept$$

Estimateur Lasso

Problèmes

Données concentrées autour de 0



Problèmes

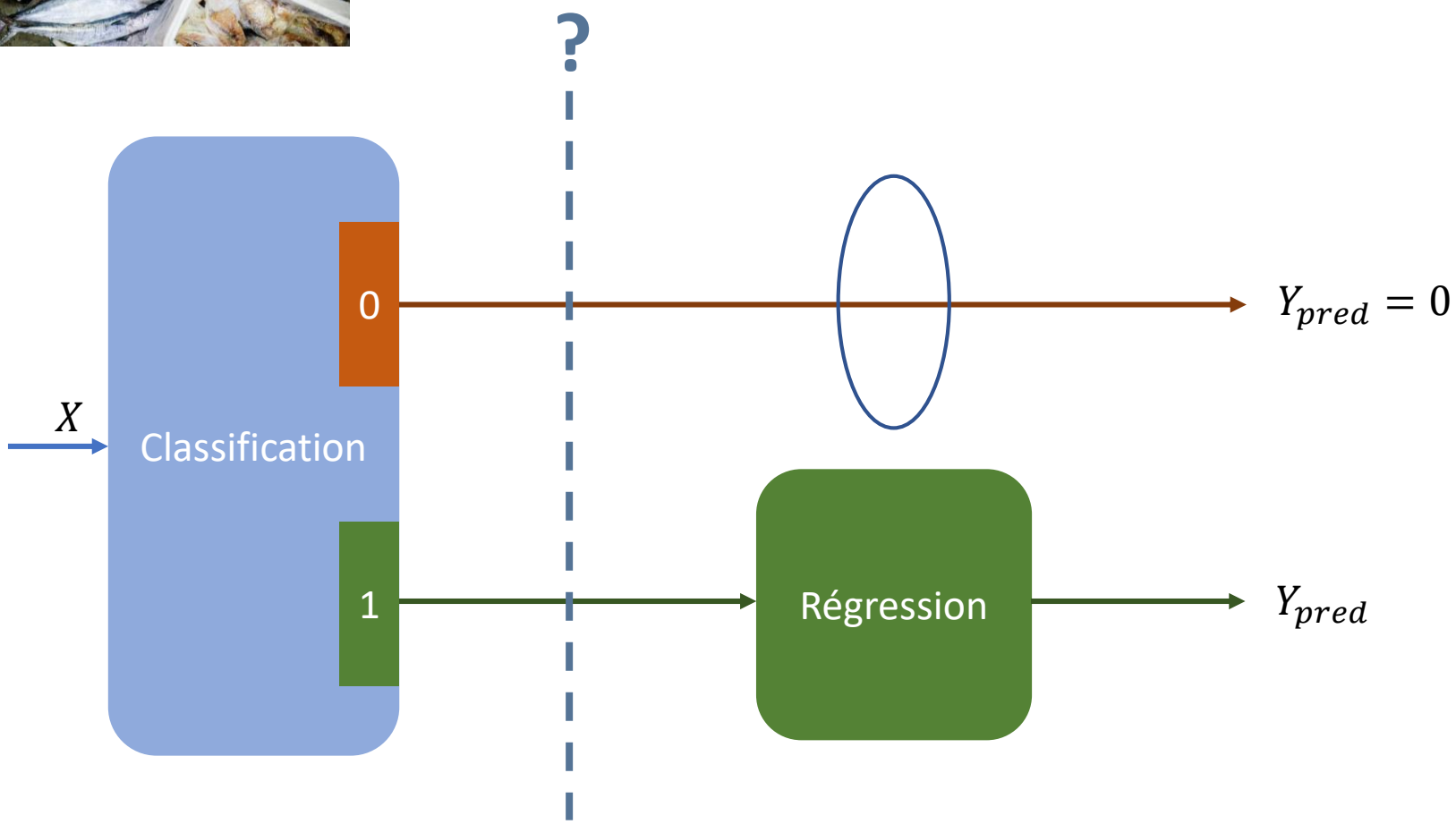
Prédiction peu satisfaisante

- La prédiction de la moyenne est souvent meilleure.
- L'estimation de présence/absence révèle un modèle inadapté.

	Donnée $= 0$	Donnée > 0
Estimation $= 0$	19802	4835
Estimation > 0	17898	543

Décomposition du problème

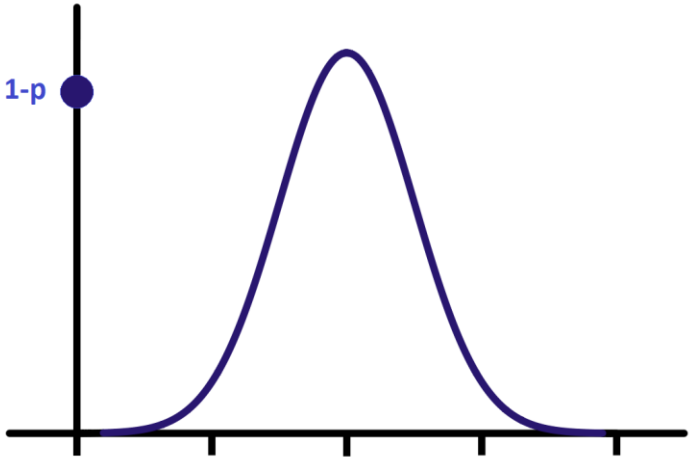
Idée de résolution



Modèle à obstacle (hurdle model)

Densité

$$f(y) = \begin{cases} 1-p & \text{si } y = 0 \\ p \times g(y) & \text{si } y \neq 0 \end{cases}$$



Introduction d'une variable binaire

- Si on pose $V = \mathbb{1}_{Y \neq 0}$
- Alors $V \sim \mathcal{B}(1, p)$

$$\begin{cases} Y|V = 0 & \sim & \delta_0, \\ Y|V = 1 & \sim & \mathcal{G}. \end{cases}$$

Décomposition de la vraisemblance

Si on estime *V* et *Y* avec des paramètres différents *γ* et *β*, la vraisemblance complète se décompose :

$$l(Y, V|X, Z; \beta, \gamma) = \underbrace{l_1(V|Z; \gamma)}_{\text{Vraisemblance typique classification binaire}} + \underbrace{l_2(Y^{(J)}|X^{(J)}; \beta)}_{\text{Vraisemblance typique régression sur le support de } Y}$$

Classification binaire

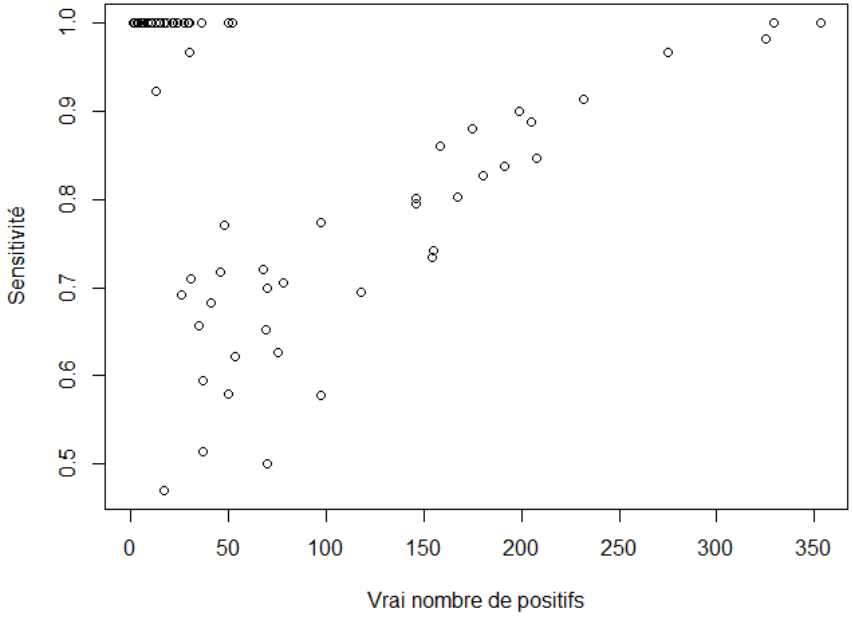
Détection de la présence de l’espèce dans les rejets

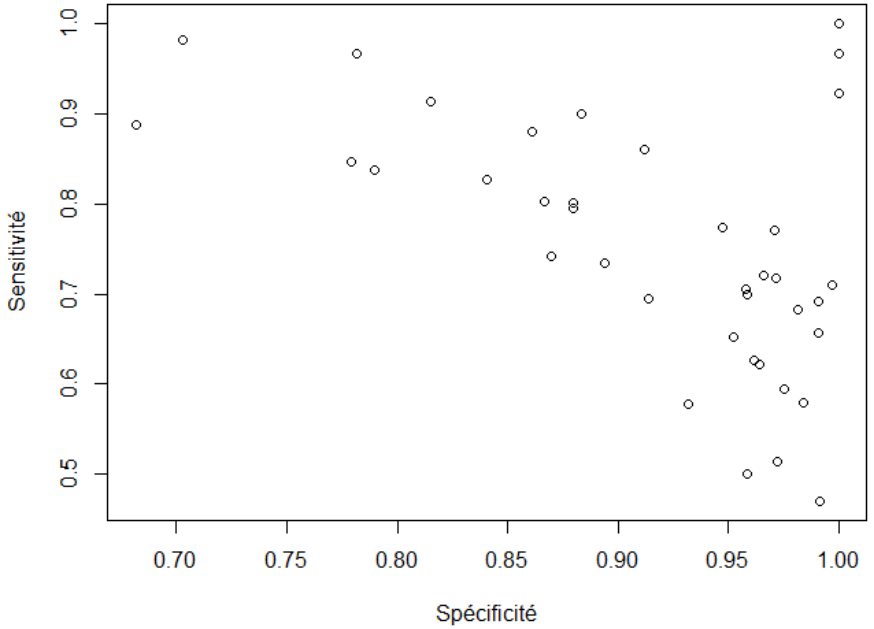
- Exemple : régression logistique

Analyse des taux de Faux Positifs et Faux Négatifs

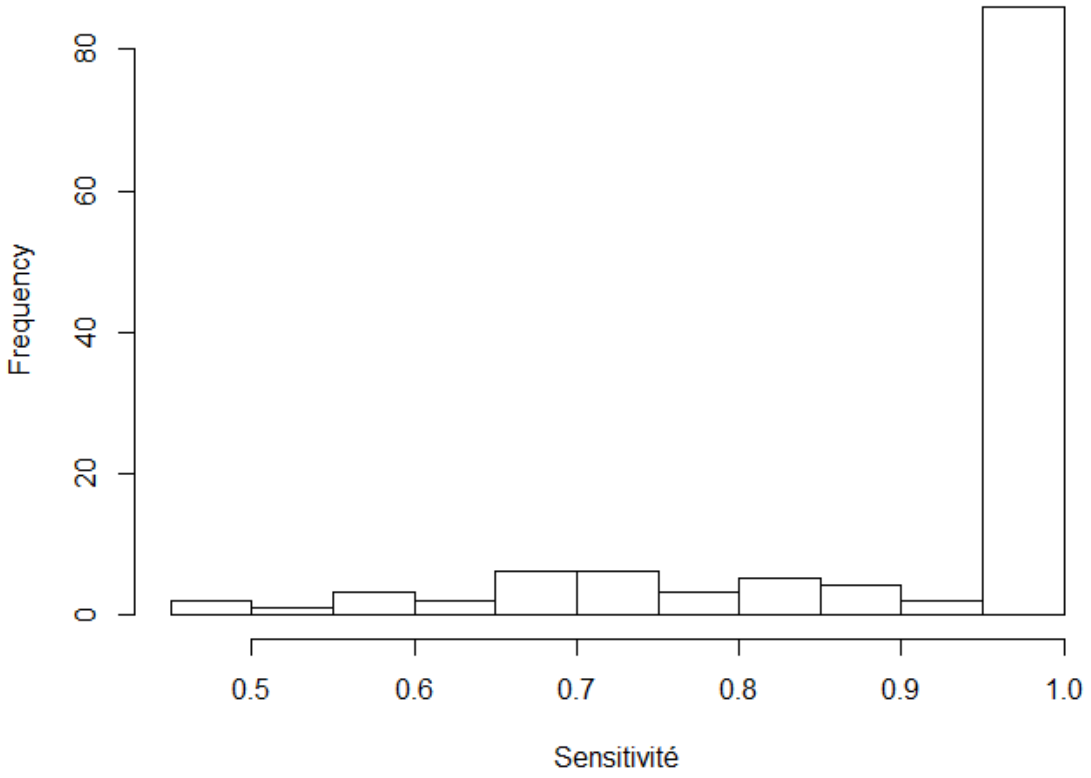
- Réglage du seuil de décision
- Besoin d’une bonne performance

Classification binaire





Classification binaire



Régression

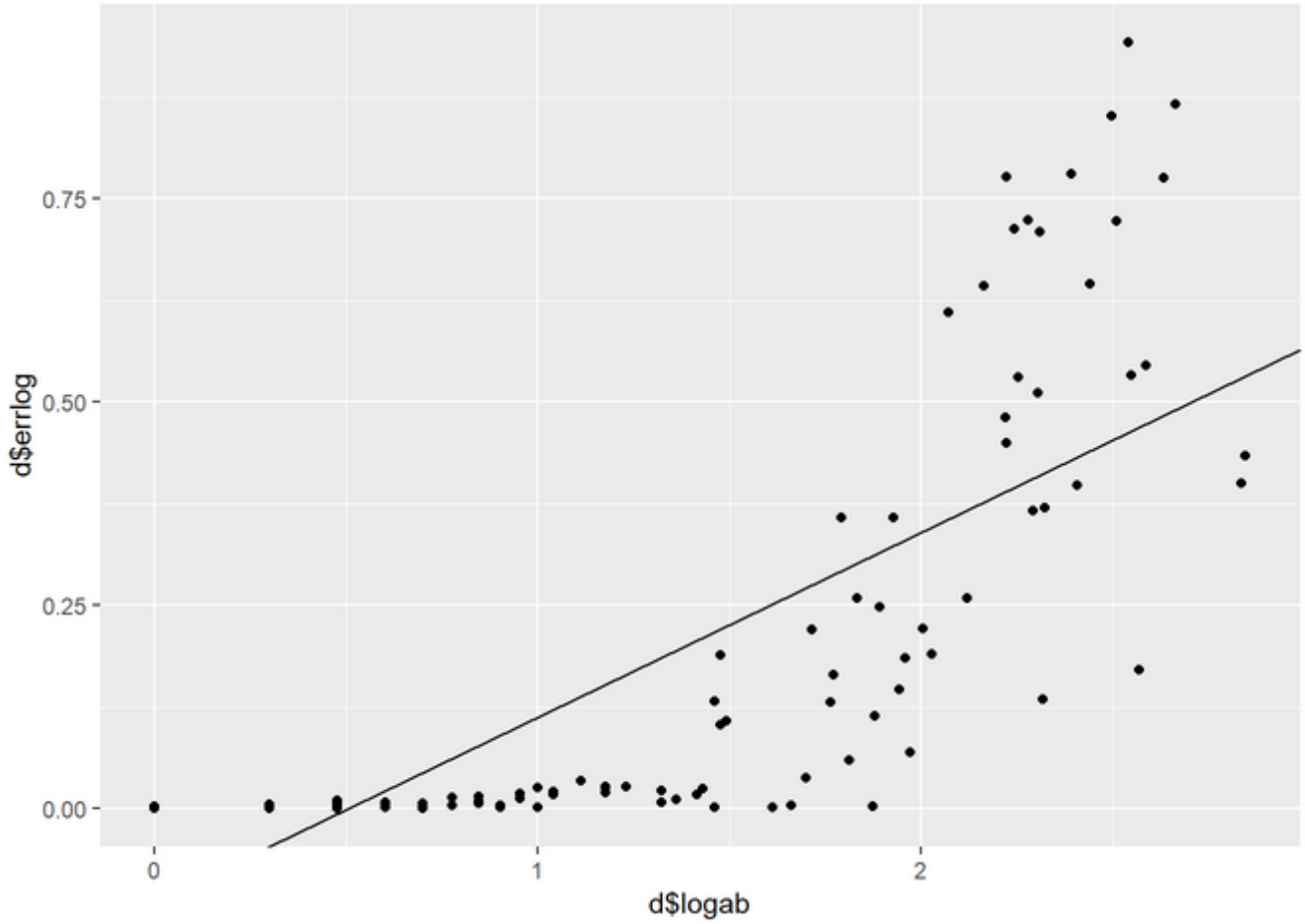
Modèles linéaires pénalisés

- L_0 : Best subset (impossible)
- Forward-Stepwise – critère AIC
- L_1 : LASSO
- Elastic-Net

Arbres de décision – forêt aléatoire

Résultats

Modèle « calque » - erreur selon abondance



Résultats

Analyse

Efficacité du modèle

- Modèle « calque »

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000000 0.002737 0.013806 0.160918 0.220786 0.941915
```

- Moyenne naïve

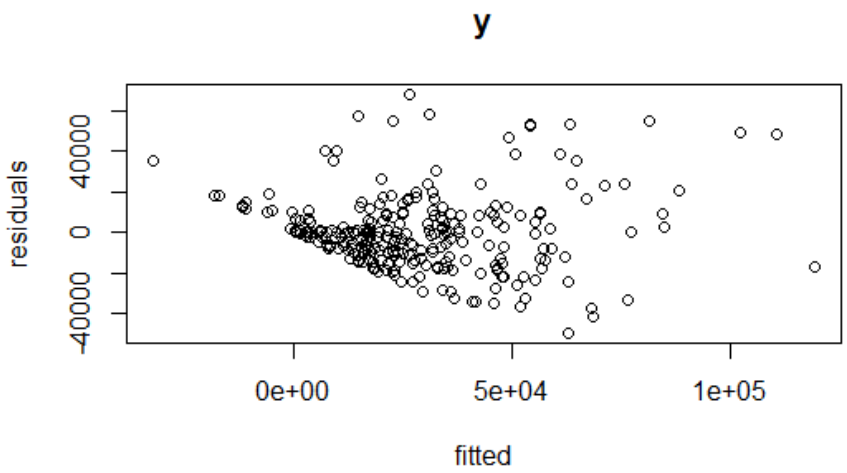
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.002023 0.834242 1.384862 1.398584 2.058610 2.827549
```

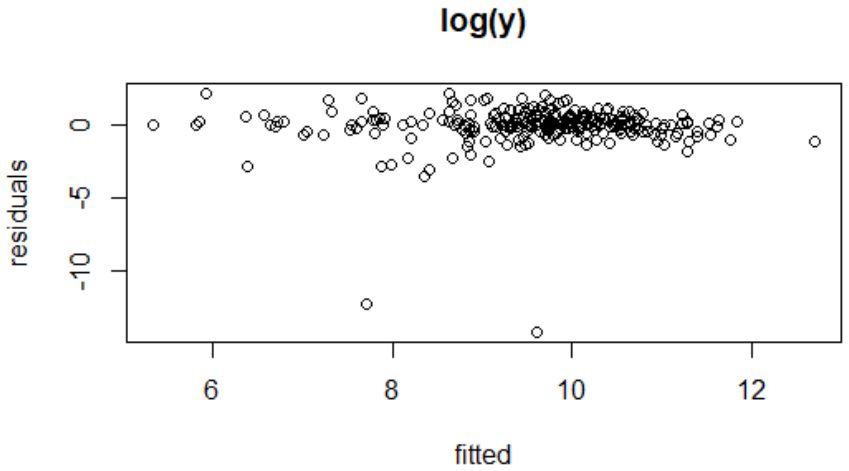
Hétéroscédasticité

Perspectives

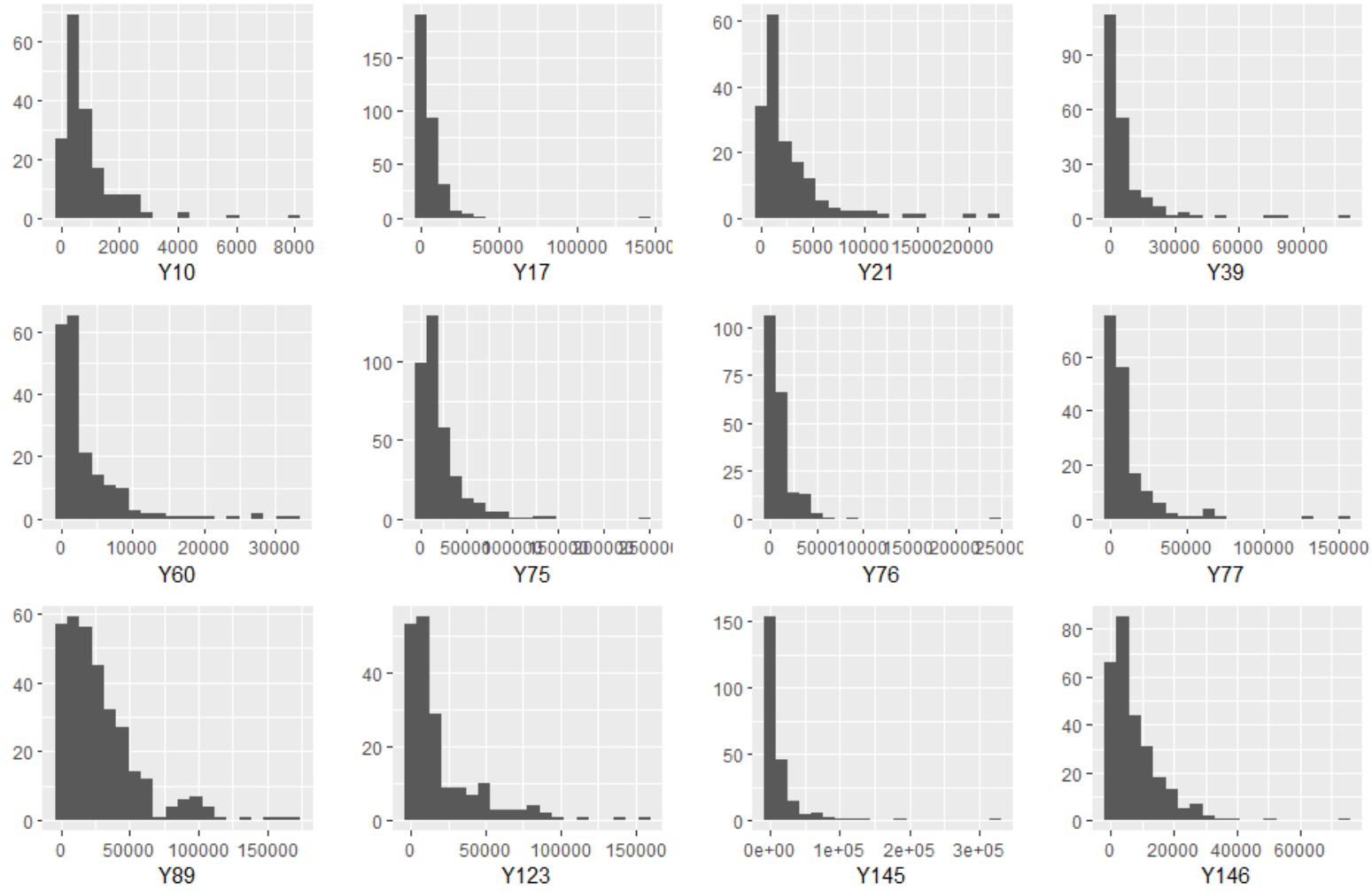
Hétéroscédasticité

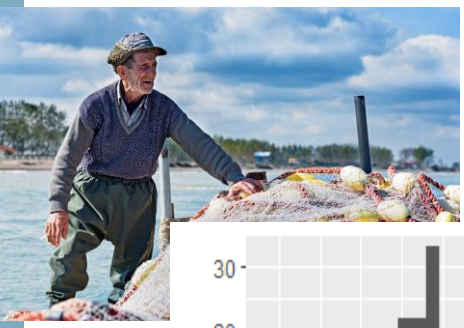
Répartition des résidus



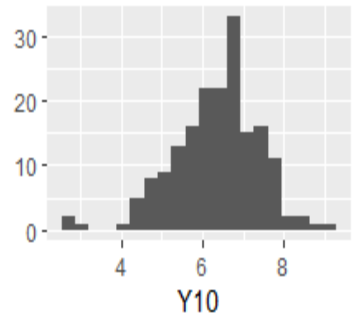


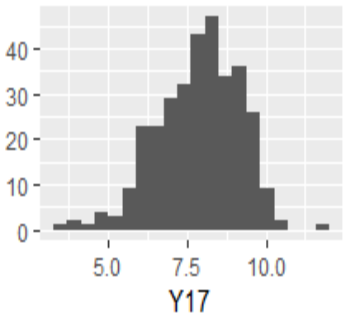
Transformation des données

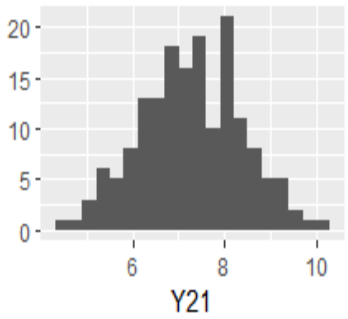


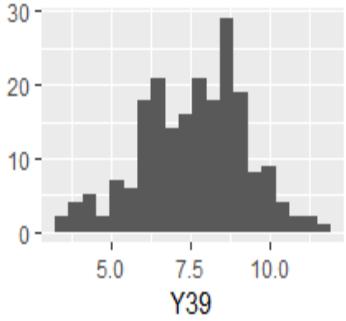


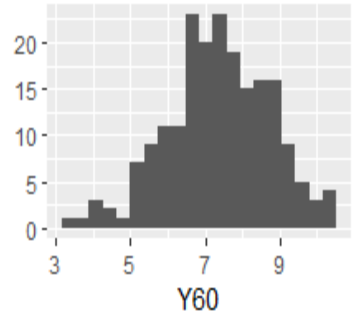
Transformation des données

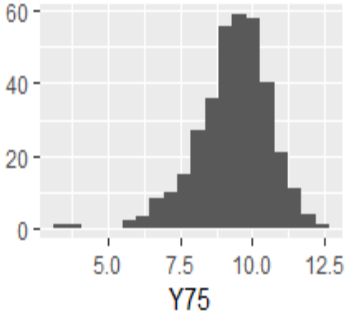


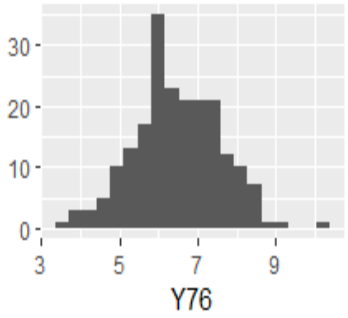


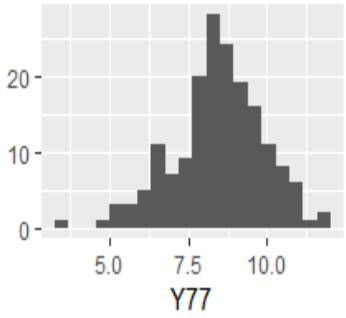


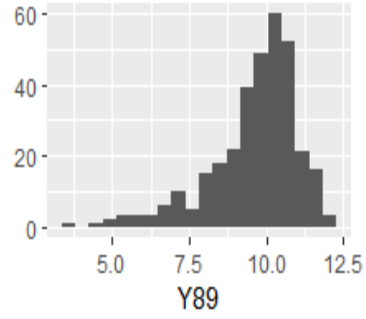


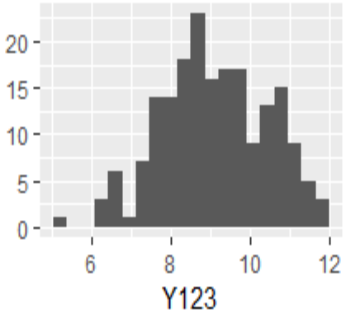


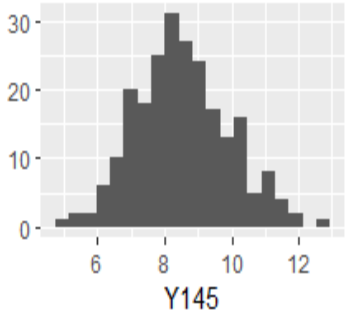


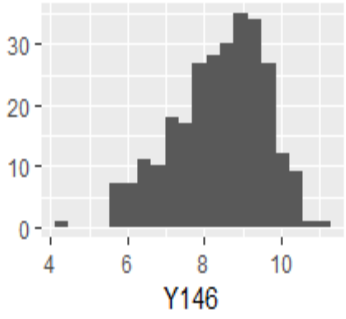












Clustering

Agrégation des cofacteurs

- Mois/géographie

Agrégation des quantités débarquées (covariables)

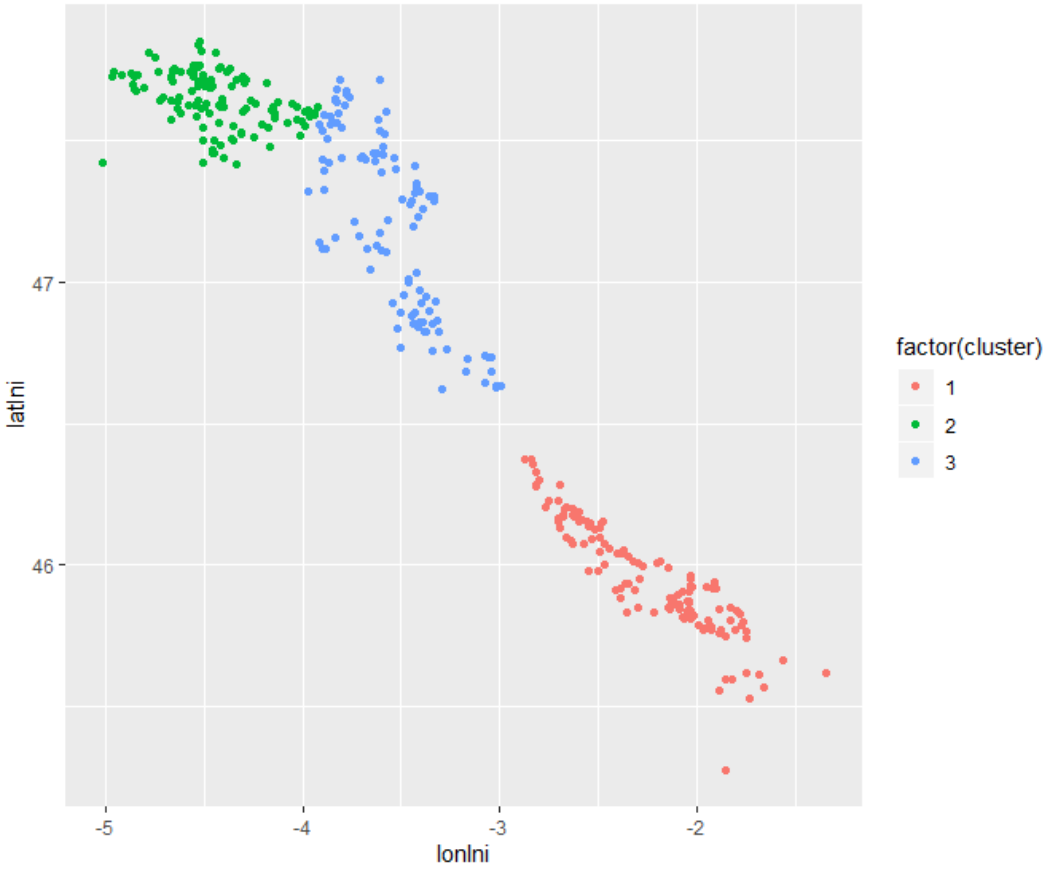
- Sélection de variables selon les clusters
- Group-Lasso

Agrégation des données de rejet

- Regrouper les données pour les régressions

Clustering

Agrégation géographique des observations



Autres approches

Validation croisée

- Problème d'équilibrage des ensembles (cofacteurs)

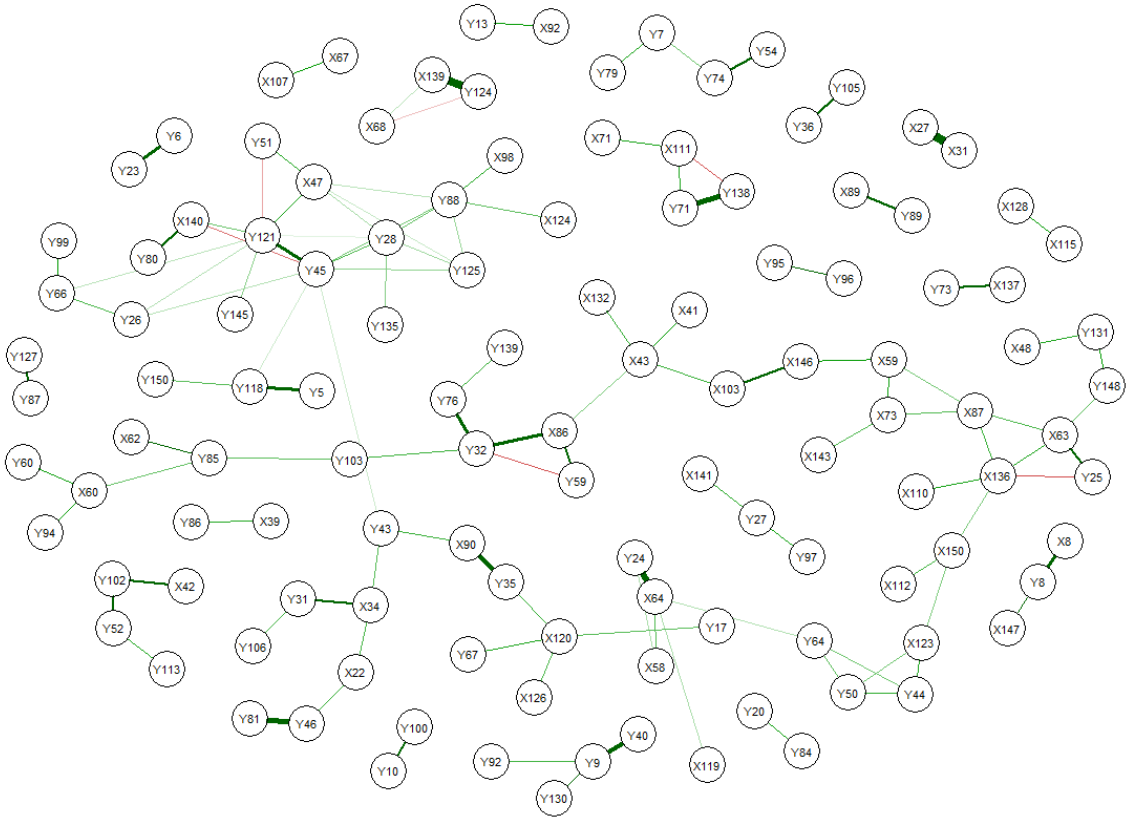
Autres méthodes de classification

- SVM
- Modèle auto-logistique

Modèles graphiques

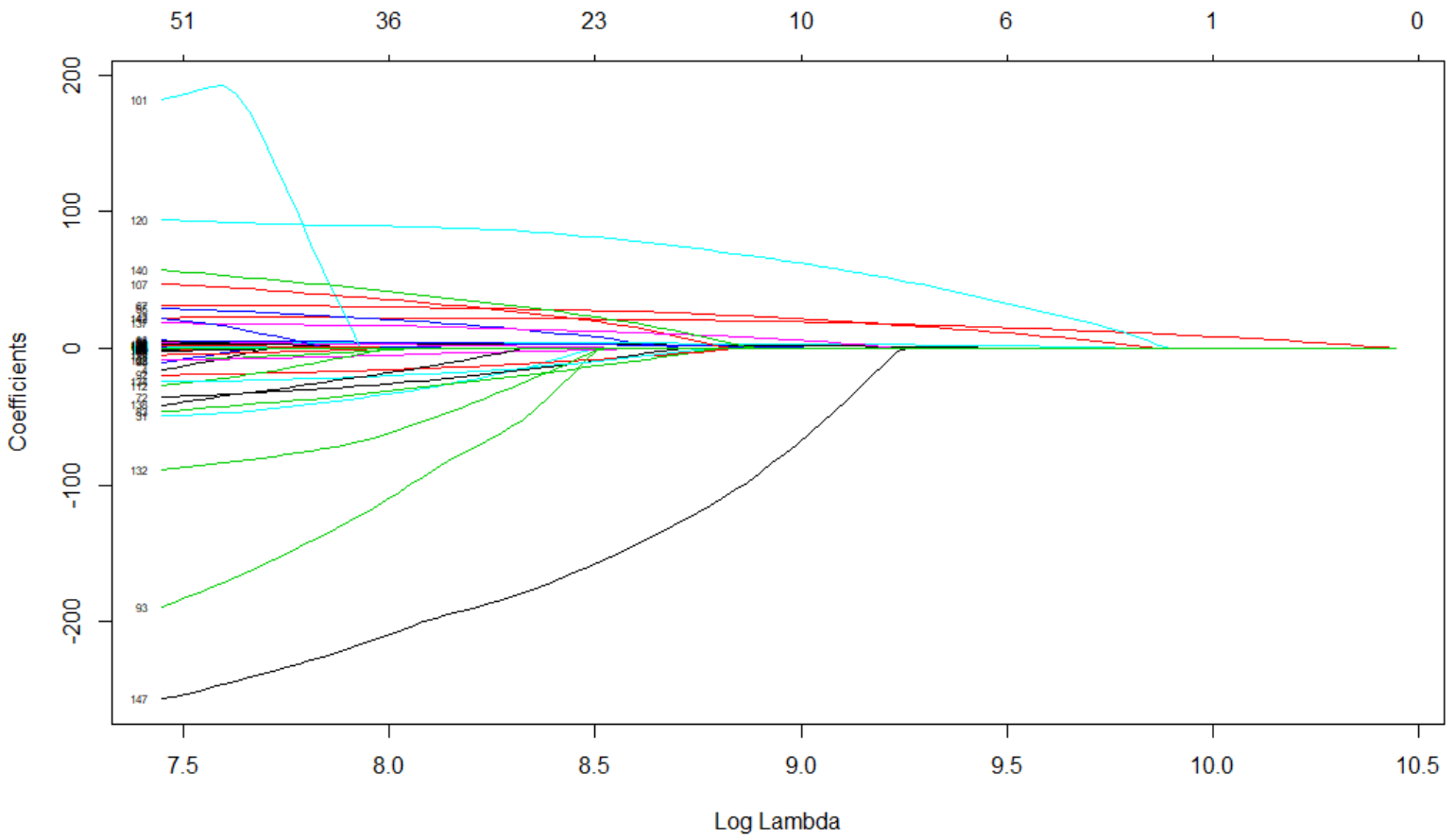
Estimation de covariance

- Idée : retrouver une structure sous-jacente
- Peut permettre la sélection de variables

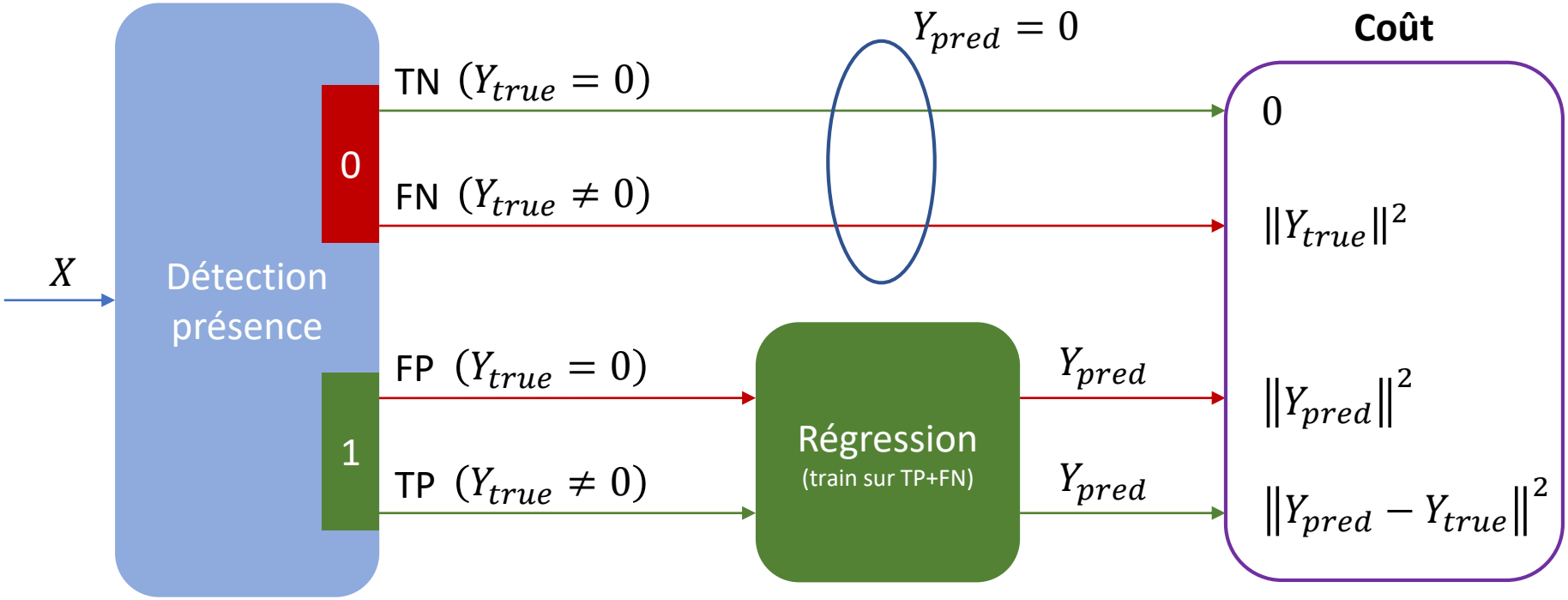



Merci !

Annexe : LASSO

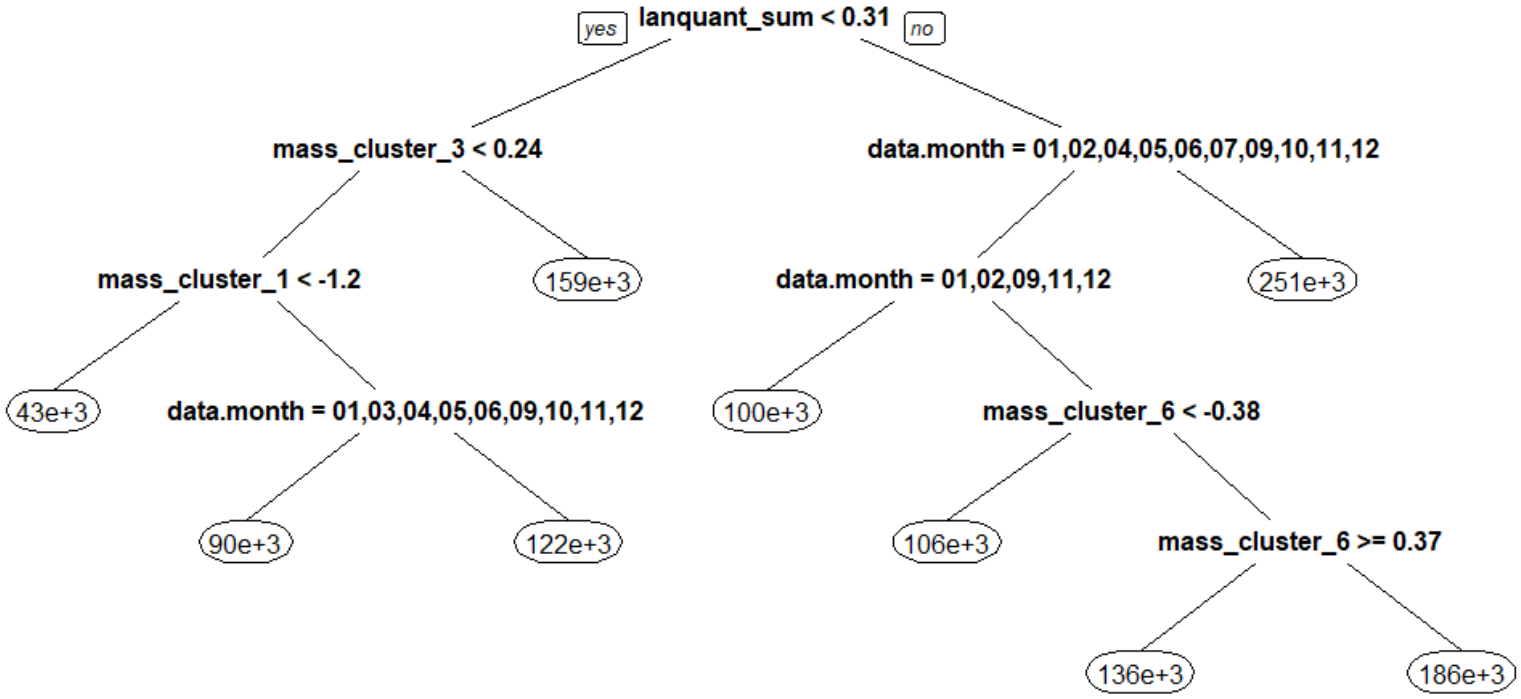


Annexe : type de coût



Annexe Régression

Arbres de décision – forêt aléatoire



Annexe : décomposition de la vraisemblance

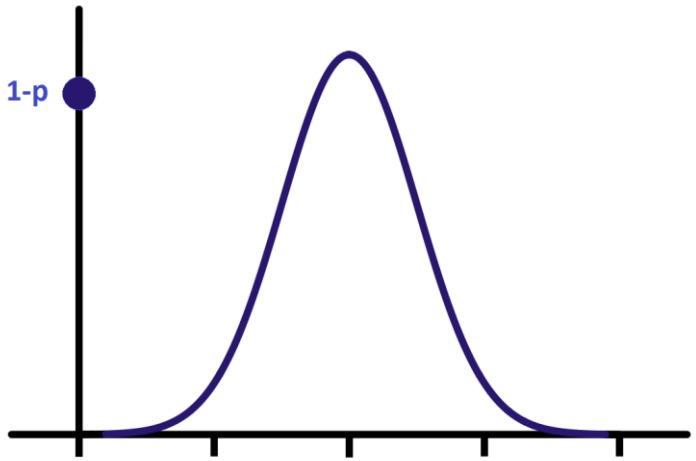
$$
\begin{aligned}
L(Y^{(i)}, V^{(i)}) &= f(Y^{(i)}|V^{(i)}=0)\mathbb{P}(V^{(i)}=0) + f(Y^{(i)}|V^{(i)}=1)\mathbb{P}(V^{(i)}=1) \\
&= \mathbb{1}_{V^{(i)}=0} \times (1-p) + g(Y^{(i)})\mathbb{1}_{V^{(i)}=1} \times p
\end{aligned}
$$

$$
\begin{aligned}
l(Y,V) &= \sum_{i=1}^{n} l(Y^{(i)}, V^{(i)}) \\
&= \sum_{i:V^{(i)}=0} \log(1-p) + \sum_{i:V^{(i)}=1} [\log p + \log g(Y^{(i)})]
\end{aligned}
$$

$$
\begin{aligned}
l(Y,V|X,Z;\beta,\gamma) &= \sum_{i:V^{(i)}=0} \log(1-p(Z^{(i)};\gamma)) + \sum_{i:V^{(i)}=1} \log p(Z^{(i)};\gamma) \\
&\quad + \sum_{i:Y^{(i)}\neq 0} \log g(Y^{(i)}|X^{(i)};\beta) \\
&= \sum_{i=1}^{n} \left\{(1-V^{(i)})\log(1-p(Z^{(i)};\gamma)) + V^{(i)} \log p(Z^{(i)};\gamma)\right\} \\
&\quad + \sum_{i\in J} \log g(Y^{(i)}|X^{(i)};\beta) \\
&= l_1(V|Z;\gamma) + l_2(Y^{(J)}|X^{(J)};\beta)
\end{aligned}
$$

Modèle à obstacle (hurdle model)

$$f(y) = \begin{cases} 1-p & \text{si } y = 0 \\ p \times g(y) & \text{si } y \neq 0 \end{cases}$$



- Si on pose $V = \mathbb{1}_{Y \neq 0}$
- Alors : $V \sim \mathcal{B}(1,p)$

$$\begin{cases} Y|V=0 & \sim & \delta_0, \\ Y|V=1 & \sim & \mathcal{G}. \end{cases}$$

$V \sim \mathcal{B}(1,p)$ $\quad V = \mathbb{I}_{Y \neq 0}$

$$\begin{cases} Y \mid V = 0 \sim \delta_0 \\ Y \mid V = 1 \sim \mathcal{G} \end{cases}$$