Master 2 MathSV 2019-2020

Rapport de Projet

Modélisation Probabiliste et Statistique de Stratégies de Pêche

20 mars 2020

Étudiants :

Tâm LE MINH
Sylvain MOINARD

Encadrants :

Sophie DONNET
Pierre BARBILLON
Stéphanie MAHÉVAS

Mots clefs : Rejets de pêche; Excès de zéros; Régression parcimonieuse; Sélection de modèle; Forêt aléatoire ; Clustering

Abstract :

Depuis 2019, la Commission Européenne interdit aux pêcheurs de rejeter à la mer les captures de poissons sous quota. Cette pratique consistant à trier les captures sur le bateau afin de ne conserver que la partie destinée à être débarquée met à mal l'évaluation des biomasses de poisson indispensables à l'établissement des quotas de pêche puisque la quantité de capture est sous-estimée. Dans la perspective d'une reprise historique des captures et à partir des données collectées par l'action ObsMer, nous avons construit des modèles de prédiction des rejets pour 150 espèces en fonction des débarquements et de facteurs environnementaux. Les modèles de régression linéaire usuels n'ont pas été concluants car ils ne permettent pas de retrouver les nombreux zéros de rejets que comportent les données. Un modèle à deux étapes avec détection de présence puis, le cas échéant, prédiction des valeurs positives s'est avéré bien plus performant. Par ailleurs, une approche du problème par des arbres de régression a également fourni des résultats prometteurs. Enfin, nous proposons des pistes pour prolonger ces travaux, notamment en agrégeant les espèces rares pour améliorer la robustesse des modèles.

Table des matières

Chapitre 1

Introduction

1 Contexte

En 2015, la Politique Commune de la Pêche de l'Union Européenne a pris une mesure interdisant les rejets des poissons pêchés en mer, avec une entrée en vigueur s'étendant jusqu'en 2019 (Commission Européenne, 2013). De fait, les pêcheurs pratiquait ces rejets dans leurs procédures courantes : les poissons étaient triés à bord, et ceux ne répondant pas aux critères commerciaux (trop petits, hors quota...) étaient rejetés à la mer souvent morts ou blessés et n'étaient donc pas pris en compte dans les quotas de pêche ni dans les suivis de la biodiversité. Désormais, tout le produit de la pêche doit être débarqué et comptabilisé dans les quotas. Cela a pour but de pérenniser les stocks de poissons et contraindre les pêcheurs à être plus sélectifs (Catchpole et al., 2017).

L'industrie halieutique a en effet des antécédents en termes de surpêche avec un impact désastreux sur la biodiversité. Comme on le voit sur la figure 1.1, la morue a été pêchée de plus en plus intensément au large de Terre-Neuve en Atlantique Nord. au cours du XXe siècle avec un pic en 1968 (810 000 tonnes pêche). S'en est suivi un déclin très rapide menant à la quasi-disparition de ce stock, jusqu'à l'interdiction de pêche en 1992 (avec une petite période de reprise entre 1998 et 2003).

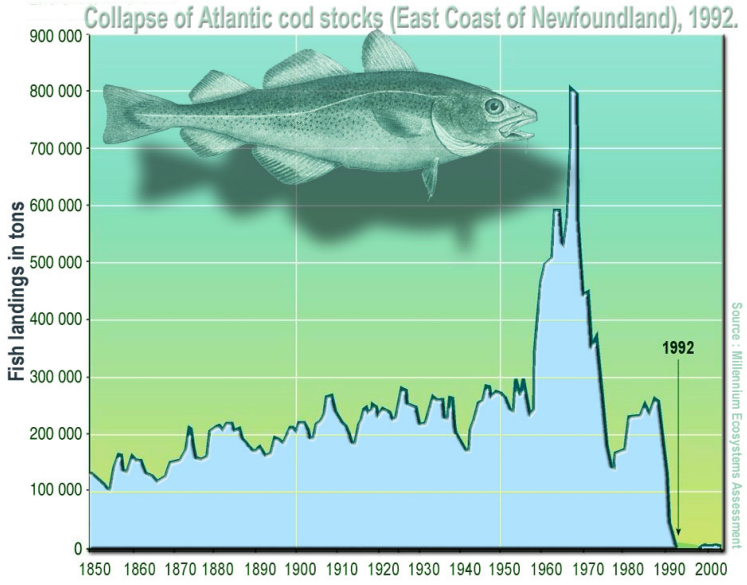

FIGURE 1.1 – Quantité de morue pêchée en fonction de l'année. Source : Wikipédia.

Depuis les années 2000, les agences de pêche européennes collectent des données dans le cadre de l'action ObsMer. Des observateurs, comme ceux de l'Ifremer en France, rejoignent des bateaux de pêche et comptabilisent les prises commerciales et les rejets (ou tout du moins en mesurent des échantillons). Notre projet a pour but d'explorer les relations entre ces quantités afin de construire un modèle de prédiction des rejets à partir des débarquements. En particulier, nous nous concentrerons sur un métier de pêche : le "chalutier langoustinier".

2 Données

2.1 Présentation

Les observations à bord des bateaux de pêche constituent un échantillonnage incomplet de l'activité de pêche car celles-ci sont trop coûteuses, et l'accueil d'observateurs à bord par les pêcheurs professionnels se fait sur la base du volontariat.

Malgré ces limites, l'échantillonnage est structuré de manière à refléter au mieux la réalité de la pêche européenne. Les observateurs sont déployés de sorte à effectuer des relevés représentatifs. Par exemple, en 2016 en France, 11% des navires ont été observés et en moyenne 86% des jours de pêche ont été pris en compte (Cornou et al., 2017). On peut noter que les estimations des quantités réellement débarquées peuvent être complétées par les informations commerciales disponibles, elles, en grande quantité. Le protocole suivi par les observateurs est standardisé quant au tri et à la pesée des captures, et les informations relatives à l'opération de pêche sont conservées : date, position...

En tout, on dispose de 362 observations d'opération de pêche associées aux masses des débarquements et des rejets de 150 espèces de poisson pour un métier de pêche et sur la période 2013-2017. Ces opérations ont celles qui ont été validées de sorte à n'avoir aucune donnée manquante. Cependant, une certaine imprécision est admise sur les quantités, du fait de la difficulté d'identifier les espèces ou de peser les poissons dans des conditions de mer variables.

Ainsi, plusieurs conséquences sont à prévoir. Le faible nombre de données peut complexifier notre problème d'inférence et inciter à la vigilance lors de l'interprétation des résultats. En revanche, la qualité des données à disposition facilite grandement l'analyse. Nous présentons ensuite une exploration quantitative des données afin d'en donner une première vue d'ensemble.

2.2 Exploration

Il s'avère que la répartition spatiale et temporelle des observations influence les captures de manière significative. Les espèces présentes et donc pêchées varient d'une région à l'autre de même que les captures dépendent de l'année et de la saison. Les observations sont aussi plus ou moins fréquentes et donc le niveau d'échantillonnage n'est pas constant.

Les observations ont été effectuées dans une portion de la façade Atlantique française correspondant approximativement à l'arc de côte qui part de la pointe bretonne et qui redescend jusqu'à l'estuaire de la Gironde, soit la moitié Nord du Golfe de Gascogne dont l'étendue est modérée (figure 1.2). Il y a bien des variations spatiales de l'abondance des espèces mais globalement chacune est présente dans toute la zone.

La distribution temporelle des observations induit de plus grandes disparités, comme on le constate sur la figure 1.3. La période d'observation court de Janvier 2013 à Décembre 2017. La répartition mensuelle révèle une concentration des observations en été avec un pic à 63 en Juin et une période creuse en hiver (seulement 7 en Janvier). Le nombre annuel des observations varie entre 61 (en 2017) et 84 (en 2016).

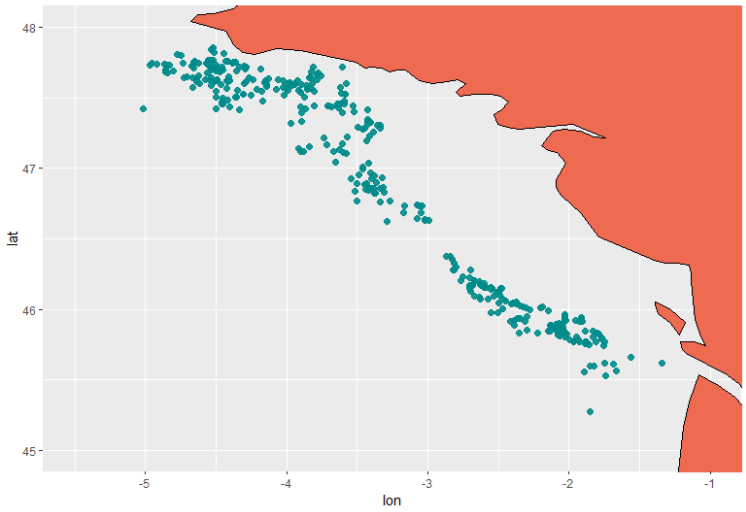

FIGURE 1.2 – Répartition spatiale des observations sur la partie Nord du golfe de Gascogne.

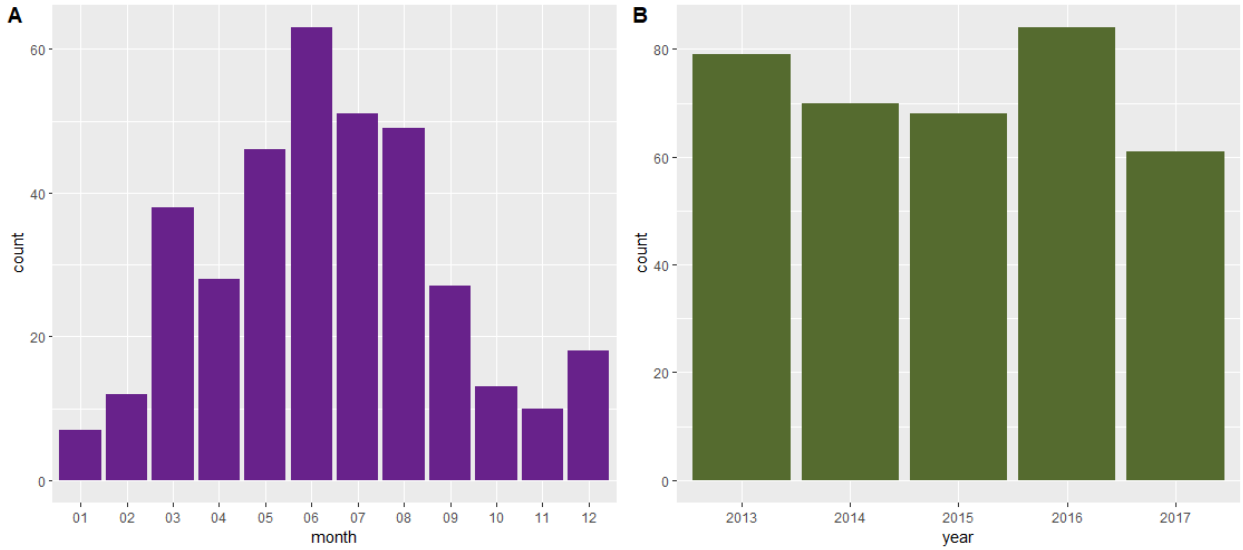

FIGURE 1.3 – Répartition temporelle des observations en fonction du mois de l'année (A) et de l'année (B). Si le nombre d'observations est relativement constant suivant les années, ce n'est pas le cas selon la saison.

Ce déséquilibre complique l'inférence. Il n'est pas raisonnable de regrouper les données par mois et par année de manière simultanée : on n'obtiendrait par exemple qu'une seule observation en Janvier 2014. Pour limiter la dispersion des données, nous avons donc choisi d'ignorer l'année d'observation pour se concentrer sur le mois, en faisant l'hypothèse que les années se ressemblent.

Les 150 espèces étudiées, quant à elles, varient beaucoup en abondance capturée. La figure 1.4 met en valeur la relation entre les quantités rejetées et débarquées et les fréquences de capture. Bien que les deux soient corrélées, considérer les unes ou les autres dans un *clustering K-means* mène à des regroupements d'espèces différents. Ces deux dimensions doivent donc être pris en compte. On peut voir leur relation sur les figures A.1 et A.2. On peut déjà augurer que les espèces rarement rejetées seront plus difficiles à estimer et on verra qu'il est préférable de prédire une absence systématique de certaines espèces. Les débarquements rares, n'ont que peu ou pas d'impact dans la prédiction des autres quantités et seront dans les faits peu utilisés dans les modèles.

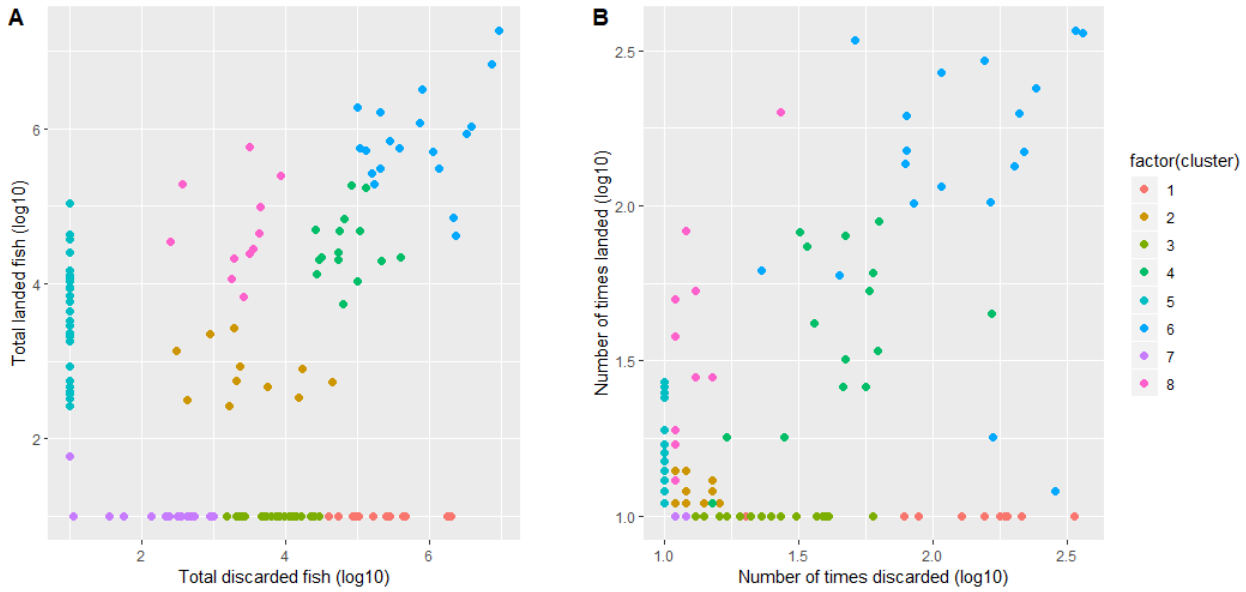

FIGURE 1.4 – Distribution des espèces en fonction de leur abondance dans les débarquements et les rejets. Chaque espèce est représentée par un point. Les sommes sur toutes les observations des quantités débarquées et rejetées sont utilisées pour réaliser un clustering (*K-means*) représenté sur la figure (A). Dans la figure (B), les espèces sont placées selon le nombre d'occurrences dans les captures tout en conservant les clusters précédents. On remarque que ceux-ci se mélangent.

Rappelons enfin qu'il sera important de prendre en compte les enjeux spécifiques de l'industrie halieutique et des suivis de biodiversité pour interpréter les résultats, qui sont complémentaires de l'analyse mathématique du problème.

3 Encodage

3.1 Observations

On pose $n = 362$ le nombre d'observations ou opérations de pêche et on classe les $K := 150$ espèces présentes dans le jeu de données par ordre alphabétique. $X_k^{(i)}$ et $Y_k^{(i)}$ correspondent aux quantités débarquée et rejetée de l'espèce k pour l'opération i.

3.2 Données temporelles

Dans le jeu de données, le mois de l'opération de pêche est codée par une variable factorielle M à 12 niveaux. Sous cette forme, elle est encodée dans les différentes régressions sous **R** par 12 variables binaires $(M_1, ..., M_{12})$ en *one-hot* : toutes nulles sauf une égale à 1 désignant le mois en question. Cette approche occulte la notion de durée entre les observations. Nous avons fait le choix de redéfinir le mois par une variable périodique pour assurer la continuité mensuelle (entre décembre de l'année y et janvier de l'année y+1) de cette information. On utilise deux variables quantitatives pour représenter chaque mois $M \in \{1, ..., 12\}$ sur le cercle trigonométrique :

$$\begin{cases} \mathrm{X}_{cos}(M) &= \cos(\frac{\pi}{6}M) \\ \mathrm{X}_{sin}(M) &= \sin(\frac{\pi}{6}M) \end{cases}. \tag{1.1}$$

Les figures A.4 et A.3 donnent une idée de la variabilité mensuelle globale des rejets et débarquements. En particulier, il y a une vraie variation des quantités rejetées à chaque opération au cours de l'année ainsi que des espèces présentes dans les captures.

3.3 Données spatiales

On dispose de deux types d'information spatiale : des coordonnées latitude-longitude et le rectangle de pêche. Ce dernier est une donnée factorielle correspondant à des zones géographiques officiellement délimitées par un maillage de 1 degré de longitude et 1/2 degré de latitude. Les coordonnées, quantitatives, sont en revanche peu précises, en particulier quant au moment de leur relevé. On va donc combiner ces deux informations. Pour chacun des L rectangles de pêche existants, on calcule la moyenne des coordonnées des opérations qui y ont eu lieu sur l'ensemble de la période 2013-2017, afin de remplacer le rectangle par les coordonnées de son centroïde, selon la formule 1.2. Cela est pertinent sous l'hypothèse que deux opérations sont d'autant plus corrélées qu'elles sont proches spatialement.

Avec $R^{(i)} \in \{1, .., L\}$ le rectangle de pêche de l'observation i, il y a donc $n_r = \sum_{i=1}^{n} \mathbb{1}_{R^{(i)}=r}$ observations dans le rectangle r. Pour $(\phi^{(i)}, \lambda^{(i)}) \in [-90, 90] \times [-180, 180]$ les coordonnées lat-lon en degrés de l'opération i, on a pour tout rectangle $r \in \{1, .., L\}$:

$$\begin{cases} \mathrm{X}_{lat}(r) &= n_r^{-1} \sum_{i=1}^{n} \mathbb{1}_{R^{(i)}=r} \phi^{(i)} \\ \mathrm{X}_{lon}(r) &= n_r^{-1} \sum_{i=1}^{n} \mathbb{1}_{R^{(i)}=r} \lambda^{(i)} \end{cases} . \tag{1.2}$$

Les figures A.6 et A.5 donnent une idée de la variabilité spatiale globale des rejets et débarquements. On voit que les quantités de poissons varient d'un rectangle de pêche à un autre, ainsi que la diversité des espèces.

3.4 Variables prédictives, variables à prédire

On rassemble les variables encodées que l'on a construit dans cette section pour construire les vecteurs des variables prédictives (les débarquements) et des variables à prédire (les rejets).

Pour la k-ième espèce, $\mathbf{X}_k$ et $\mathbf{Y}_k \in \mathbb{R}^n$ représentent les quantités de poissons débarquées et rejetées à chaque opération. De manière similaire, on note $X^{(i)}$ le vecteur des réalisations des variables prédictives et $Y^{(i)}$ celui des variables à prédire pour l'opération i :

$$X^{(i)} = \begin{pmatrix} \mathrm{X}_1^{(i)} \\ \dots \\ \mathrm{X}_K^{(i)} \\ \mathrm{X}_{cos}^{(i)} \\ \mathrm{X}_{sin}^{(i)} \\ \mathrm{X}_{lat}^{(i)} \\ \mathrm{X}_{lon}^{(i)} \end{pmatrix}, \qquad Y^{(i)} = \begin{pmatrix} \mathrm{Y}_1^{(i)} \\ \dots \\ \mathrm{Y}_K^{(i)} \end{pmatrix},$$

et on construit les vecteurs $\mathbf{X}_j$ et $\mathbf{Y}_j$ et les matrices $\mathbf{X}$ et $\mathbf{Y}$ regroupant toutes les observations :

$$\mathbf{X} = \begin{pmatrix} \text{\_\_\_\_\_}{}^t\mathrm{X}^{(1)}\text{\_\_\_\_\_} \\ \text{\_\_\_\_\_}{}^t\mathrm{X}^{(2)}\text{\_\_\_\_\_} \\ \cdots \\ \text{\_\_\_\_\_}{}^t\mathrm{X}^{(n)}\text{\_\_\_\_\_} \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_{lon} \\ | & | & & | \end{pmatrix}, \tag{1.3}$$

$$\mathbf{Y} = \begin{pmatrix} \text{\_\_\_\_\_}{}^t\mathrm{Y}^{(1)}\text{\_\_\_\_\_} \\ \text{\_\_\_\_\_}{}^t\mathrm{Y}^{(2)}\text{\_\_\_\_\_} \\ \cdots \\ \text{\_\_\_\_\_}{}^t\mathrm{Y}^{(n)}\text{\_\_\_\_\_} \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \mathbf{Y}_1 & \mathbf{Y}_2 & \ldots & \mathbf{Y}_K \\ | & | & & | \end{pmatrix}. \tag{1.4}$$

Chapitre 2

Modèles de régression linéaire

1 Écriture générale du modèle

1.1 Mise en équation

Avant toute chose, nous avons observé la corrélation entre les valeurs des x et des y. En effet, il serait attendu un certain lien entre les rejets et les débarquements d'une même espèce donnée. Ce n'est en fait pas vraiment le cas. En premier lieu, seulement 56 des 150 espèces sont présentes à la fois dans les débarquements et les rejets. En second lieu, la matrice des corrélation rejets-débarquements révèle que ce lien est loin d'être évident (figure 2.1). Nous choisirons donc un modèle avec comme variables explicatives les données de débarquement de l'ensemble des espèces et non de l'espèce à prédire uniquement.

Nous avons tout d'abord utilisé une régression linéaire classique pour prédire les quantités rejetées pour chaque espèce à partir des quantités débarquées et des variables spatiales et temporelles. Pour chaque opération i et chaque espèce k, on cherche un modèle de la forme suivante, avec les erreurs iid $(\epsilon_k^{(i)}) \sim \mathcal{N}(0, \sigma^2)$ où $\sigma^2 > 0$ représente l'ampleur de l'écart au modèle (et dont la distribution gaussienne est une hypothèse du modèle). $\forall k \in \{1, ..., K\}, \forall i \in \{1, ..., n\}$,

$$Y_k^{(i)} = \sum_{l=1}^{K} \beta_l^k . X_k^{(i)} + \beta_{lon}^k . X_{lon}^{(i)} + \beta_{lat}^k . X_{lat}^{(i)} + \beta_{cos}^k . X_{cos}^{(i)} + \beta_{sin}^k . X_{sin}^{(i)} + \beta_k^0 + \epsilon_k^{(i)}.$$

On peut écrire le problème sous forme matricielle :

$$Y^{(i)} = \beta X^{(i)} + \beta^0 + \epsilon^{(i)}$$

selon les notations présentées précédemment et avec :

- $\beta^0 = {}^t(\beta_1^0 \dots \beta_K^0)$,
- $\epsilon^{(i)} = {}^t(\epsilon_1^{(i)} \dots \epsilon_K^{(i)})$,
- $\mathbb{1} = (1, ..., 1) \in \mathbb{R}^{1 \times n}$,
- $\beta = \begin{pmatrix} \beta_1^1 & \dots & \beta_K^1 & \beta_{cos}^1 & \beta_{sin}^1 & \beta_{lat}^1 & \beta_{lon}^1 \\ \vdots & & \dots & & & \dots & \\ \beta_1^K & \dots & \beta_K^K & \beta_{cos}^K & \beta_{sin}^K & \beta_{lat}^K & \beta_{lon}^K \end{pmatrix}$.

Ainsi, pour l'ensemble des opérations de pêche, avec $\epsilon = (\epsilon_k^{(i)})_{k,i} \in \mathbb{R}^{K \times n}$, on a :

$${}^t\mathbf{Y} = \beta . {}^t\mathbf{X} + \beta^0 . \mathbb{1} + \epsilon.$$

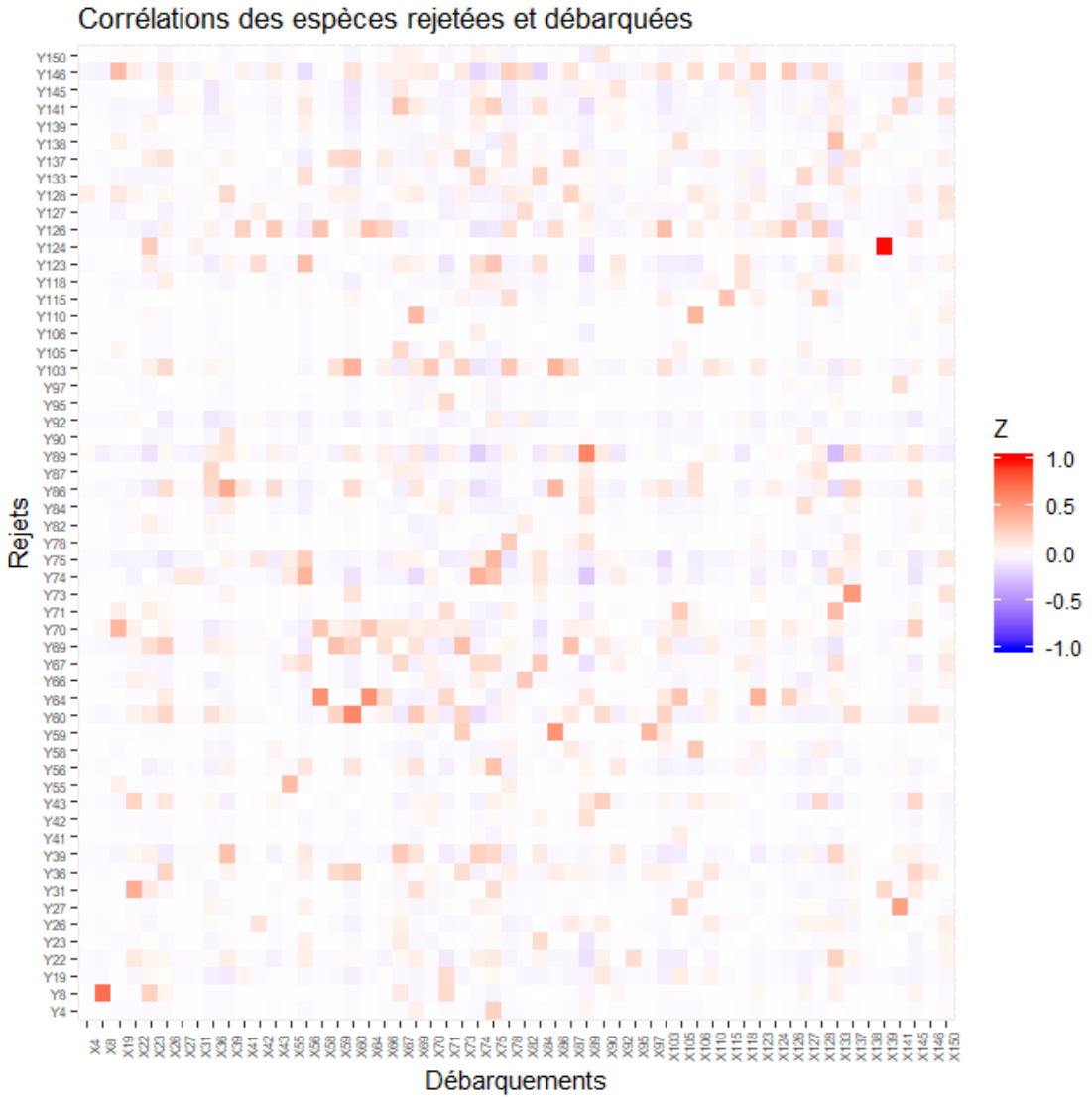

FIGURE 2.1 – Matrice de corrélation (*heatmap*) pour les 56 espèces à la fois rejetées et débarquées. On constate que les valeurs rouges et bleu foncé, qui correspondent à une corrélation élevée (en valeur absolue) sont peu nombreuses. On en retrouve quelques unes sur la diagonale principale (ie $\text{Cor}(X_k, Y_k)$ pour chaque espèce k) mais ces corrélations intraspécifiques ne sont pas évidentes en général.

L'enjeu est d'estimer les coefficients de régression β de manière à minimiser l'erreur quadratique pour chaque espèce k :

$$(\hat{\beta_k}, \hat{\beta_k^0}) = \underset{\beta_k, \beta_k^0}{\text{argmin}} ||\beta_k X^{(i)} + \beta_k^0 - Y_k^{(i)}||_2.$$

Ainsi, l'estimateur retenu est :

$$\hat{Y}_k^{(i)} = (\hat{\beta_k} X^{(i)} + \hat{\beta_k^0})_+$$

où $(.)_+$ désigne la partie positive. On refuse en effet les prédictions négatives qui n'ont pas de sens sur le plan écologique.

1.2 Dimension réelle du problème

Dans les faits, on cherche à prédire les rejets pour $\tilde{K} = 120$ espèces car 30 espèces des données n'ont aucun rejet. De plus, seulement 86 espèces sont débarquées au moins une fois. En prenant en compte les quatre variables environnementales, la dimension réelle de β est donc 120×90.

2 Sélection de modèles

Un problème apparaît directement : dans la configuration présentée précédemment, toutes les variables explicatives sont utilisées pour déterminer β, alors même qu'il est probable qu'un grand nombre d'entre elles ne soient pas statistiquement significatives. Ce modèle trop naïf ne peut donc pas être performant en termes de robustesse. Il faut s'attendre à ce que le modèle ait sur-appris à partir des données d'entraînement et que son application à des nouvelles données ne donne pas des résultats satisfaisants.

Best subset Dans l'idéal, on voudrait déterminer un *Best subset* de variables significatives qui seraient les seules à être associées à des coefficients de régression non nuls. Dans les faits, on ne sait pas déterminer ce groupe parfait car il faudrait tester la présence de chaque variable dans toutes les configurations possibles, ce qui ferait ici $2^{154} \sim 10^{46}$ possibilités. On utilise donc différents approximations pénalisées.

Critère AIC La première option est la pénalisation du modèle par le critère AIC (Critère d'Information d'Akaike), qui s'écrit pour chaque espèce k :

$$(\hat{\beta_k}, \hat{\beta_k^0}) = \underset{\beta_k, \beta_k^0}{\operatorname{argmin}}(||\beta_k X^{(i)} + \beta_k^0 - Y_k^{(i)}||_2^2 + 2d\sigma^2)$$

où d est le nombre de composantes non-nulles de β_k. On utilise pour cela une approche *stepwise* avec la fonction **step** du package **stats** de **R**, dans le mode *forward-backard*. L'algorithme démarre en prédisant la moyenne (formule "$Y_k \sim 1$"). Puis à chaque itération il teste l'ajout de chaque variable absente ou le retrait de chaque variable présente et conserve le set de variables qui mène au plus bas AIC. La taille du set varie donc de 1 en 1. L'algorithme se termine quand aucune modification des variables ne réduit l'AIC.

Régression pénalisée D'autres types d'estimateurs pénalisent les valeurs des composantes. Plusieurs sont implémentés dans le package **glmnet** (Friedman et al., 2010). Le plus usuel est l'estimateur LASSO (Least Absolute Shrinkage and Selection Operator), avec une pénalisation en norme ℓ_1, dont les paramètres de régression vérifient :

$$(\hat{\beta_k}, \hat{\beta_k^0}) = \underset{\beta_k, \beta_k^0}{\operatorname{argmin}}(||\beta_k X^{(i)} + \beta_k^0 - Y_k^{(i)}||_2^2 + \lambda|\beta_k|_1).$$

L'avantage de la pénalisation en norme ℓ_1 est qu'elle sélectionne les variables : celles qui participent peu sont neutralisées. Pour cette raison, on parle de régression parcimonieuse. Il est également commun de pénaliser les paramètres selon la norme ℓ_2 (Hoerl & Kennard, 1970) :

$$(\hat{\beta_k}, \hat{\beta_k^0}) = \underset{\beta_k, \beta_k^0}{\operatorname{argmin}}(||\beta_k X^{(i)} + \beta_k^0 - Y_k^{(i)}||_2^2 + \lambda||\beta_k||_2^2).$$

Cependant, contrairement à la norme ℓ_1, cela réduit mais ne permet pas de jeter les variables peu significatives. On pourrait considérer la combinaison des deux, qui s'appelle l'estimateur Elastic-Net (Zou & Hastie, 2005) :

$$(\hat{\beta_k}, \hat{\beta_k^0}) = \underset{\beta_k, \beta_k^0}{\operatorname{argmin}}(||\beta_k X^{(i)} + \beta_k^0 - Y_k^{(i)}||_2^2 + \lambda||\beta_k||_2^2 + \mu|\beta_k|_1).$$

Dans la section suivante, les résultats de modèle linéaire présentés sont ceux obtenus par la méthode **step** portant sur l'AIC. L'estimateur LASSO sera utilisé dans un autre chapitre de ce rapport.

3 Ensemble de test

Si on entraîne le modèle de régression pénalisée sur toutes les données disponibles, il est probable que son application à de nouvelles données ne soit pas encore assez performant car trop ajusté aux premières données, d'autant plus que ledit modèle est complexe (ce qu'on mesure ici avec le nombre de coefficients de régression non-nuls). Pour pallier cela, on peut séparer le jeu de données en deux.

On sélectionne une partie des données, par exemple 70% ($n_{train} = \lfloor 0.7n \rfloor$), qui constitue l'ensemble d'entraînement qui servira à construire le modèle. Ensuite, on teste ce modèle sur les 30% de données restantes qui composent l'ensemble de test ($n_{test} = n - n_{train}$).

Les résultats sur l'ensemble de test permettront de valider le modèle. On a ainsi un meilleur aperçu des performances réelles du modèle. Nous avons choisi de prédire systématiquement 0 pour les espèces avec moins de 5 observations de rejets dans l'ensemble d'entraînement (ie pour les espèces d'abcisse proche de 0 sur la figure 1.4 B). Il aurait été également possible de prédire la moyenne des rejets avec l'estimateur suivant, constant pour chaque espèce :

$$\hat{Y}_k^{(i)} = \frac{1}{n}\sum_{j=1}^{n} Y_k^{(j)}, \qquad \forall 1 \leq i \leq n, \quad \forall 1 \leq k \leq K.$$

Nous n'avons pas travaillé spécifiquement sur la représentativité des données. Pour les espèces rares, il serait intéressant de faire en sorte que des opérations comptant des rejets ou débarquements de ces espèces soient systématiquement sélectionnées. Cela rendrait le modèle plus robuste, car pour les espèces tangentes (par exemple avec une dizaine d'occurrences), on prédit ou non des rejets positifs selon l'aléa de composition de l'ensemble d'entraînement.

Toujours dans l'optique de réduire le surajustement, nous introduirons également plus tard des techniques de validation croisée, comme le V-fold.

4 Analyse des résultats

4.1 Métrique de comparaison

Il a fallu définir une métrique pour quantifier l'erreur de prédiction. Un choix classique pourrait être la distance euclidienne usuelle entre les valeurs prédites ($\hat{Y}_k^{(i)}$) et les valeurs réelles ($Y_k^{(i)}$) : $(\hat{Y}_k^{(i)} - Y_k^{(i)})^2$. Cependant, en prévision d'une faible précision des résultats et au vu des ordres de grandeurs très variables des données, il a été retenu une métrique logarithmique (en base 10). Pour

une espèce donnée k, cette distance est :

$$\Delta(\hat{Y}_k^{(i)}, Y_k^{(i)}) = \left| \log_{10} \left(\frac{\hat{Y}_k^{(i)} + 1}{Y_k^{(i)} + 1} \right) \right|.$$

Ainsi, le coût de l'erreur pour l'espèce k est :

$$J_k^{log} = \frac{1}{n} \sum_{i=1}^{n} \left| \log_{10} \left(\frac{\hat{Y}_k^{(i)} + 1}{Y_k^{(i)} + 1} \right) \right|.$$

J_k^{log} mesure le nombre moyen d'ordres de grandeur entre la prédiction et la valeur exacte. Pour être exact, la valeur moyenne des (J_k^{log}) est le logarithme de la moyenne géométrique de $max\left(\frac{Y_k^{(i)}}{\hat{Y}_k^{(i)}}, \frac{\hat{Y}_k^{(i)}}{Y_k^{(i)}}\right)$, puisque :

$$10^{J^{\log}} = \left(\prod_{k=1}^{K} \prod_{i=1}^{n} \frac{\max(\hat{Y}_k^{(i)} + 1, Y_k^{(i)} + 1)}{\min(\hat{Y}_k^{(i)} + 1, Y_k^{(i)} + 1)} \right)^{\frac{1}{nK}}.$$

4.2 Résultats

Pour le modèle linéaire avec sélection de variables *stepwise* sur le critère AIC, on obtient la répartition de l'erreur représentée sur la figure 2.2 pour l'ensemble d'entraînement. La forme est très comparable pour l'ensemble de test. 62 espèces dépassent le seuil des 5 rejets positifs dans la partition étudiée. Les 88 autres espèces qui ne dépassent pas ce seuil sont décrites dans la table B.1.

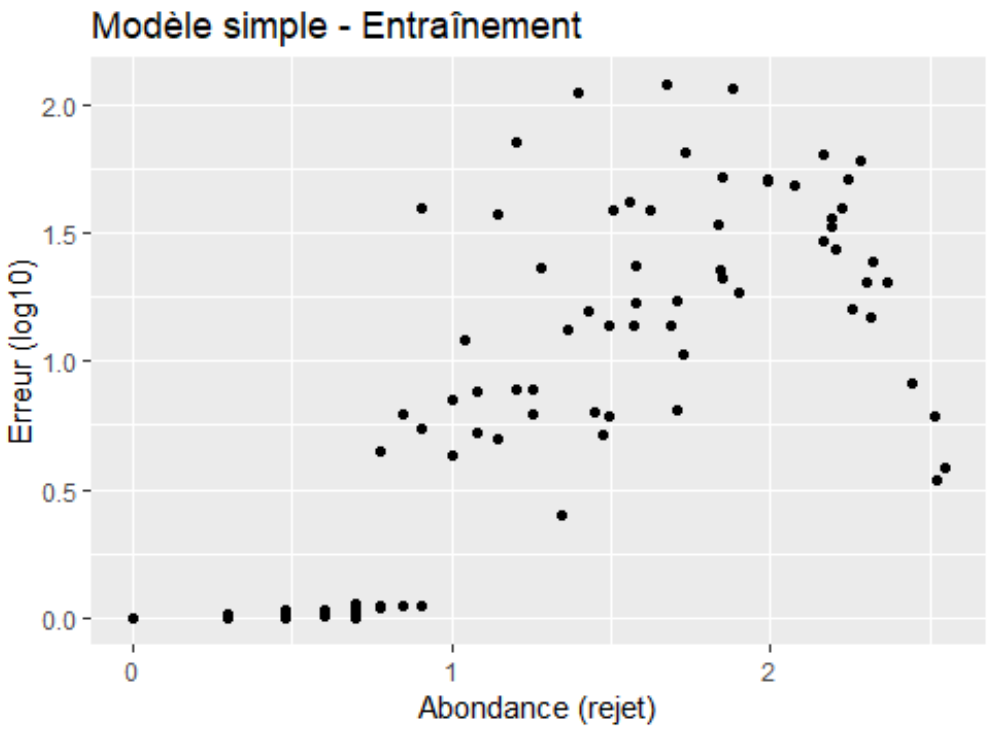

FIGURE 2.2 – Erreur J_k^{log} pour le modèle de régression linéaire simple avec sélection de variables AIC sur les données d'entraînement.

Chaque point représente une espèce, et l'abscisse est le logarithme du nombre d'opérations de pêche où l'espèce est présente dans les rejets. La prédiction systématique de 0 pour les espèces peu présentes conduit à une faible erreur tandis que celle-ci est grande pour les espèces abondamment rejetées. On constate une discontinuité au niveau du seuil des 5 rejets positifs dans l'ensemble d'entraînement : quand nous tentons des prédictions non triviales sur les espèces rares, nous n'y parvenons pas très efficacement. La répartition de l'erreur est résumée dans le tableau 2.1.

Ensemble	Médiane	Moyenne	Maximum
Entraînement	0.026	0.52	2.08
Test	0.033	0.57	2.53

TABLE 2.1 – Répartition de l'erreur J_k^{log} pour le modèle de régression linéaire simple avec sélection de variables AIC. Les 150 espèces sont prises en compte. En moyenne sur les nouvelles données, il y a un rapport $10^{0.57} \simeq 3.7$ entre les données et l'estimation.

D'autres diagnostics sont réalisables pour évaluer l'estimateur. On constate que seulement 6.8% des coefficients de régression sont non-nuls (381 sur les 62×90 possibles) : le modèle décide de n'incorporer qu'une petite partie des données disponibles.

De plus, on sait que les données comptent un grand nombre de $(Y_k^{(i)} = 0)$: le modèle reflète-t-il cette spécificité ? Si on partitionne les données selon $(Y_k^{(i)} > 0)$ contre $(Y_k^{(i)} = 0)$, on peut dresser quatre catégories de prédictions indiquées dans la table 2.2. Si on applique ce protocole à nos données, on obtient un résultat peu satisfaisant comme on le voit dans la table de confusion 2.3. Le résultat est juste dans 77% des cas. Il y a notamment un excès de Faux Positifs : le modèle actuel ne sait pas retrouver les valeurs nulles des données.

On peut calculer deux autres grandeurs pour évaluer l'efficacité de prédiction. La sensibilité correspond au taux de détection des valeurs positives : $\frac{\text{VP}}{\text{VP} + \text{FN}}$ et vaut ici 88% ; la spécificité correspond au taux de détection des valeurs nulles : $\frac{\text{VN}}{\text{VN} + \text{FP}}$ et est de 76%.

Le même protocole a été appliqué au modèle qui prédit pour chaque espèce la moyenne des quantités rejetées (0 inclus). Les résultats sont résumés dans la table 2.4 : ils sont légèrement moins bons que dans le modèle plus complexe de la régression linéaire. La figure 2.3 montre que la répartition de l'erreur est similaire.

4.3 Difficultés rencontrées

Plusieurs problèmes émergent de cette première analyse. Tout d'abord, le modèle linéaire ne semble pas significativement meilleur que la simple prédiction de la moyenne, et une faible partie de l'information disponible semble utilisée. En outre, l'abondance de 0 des données est mal détectée. Enfin, la répartition de l'erreur n'est pas uniforme selon l'abondance et les volumes considérés, comme on le voit en observant les résidus sur la figure 2.4. Il s'agit d'un phénomène d'hétéroscédasticité. L'erreur croît avec ladite abondance des rejets, et une discontinuité est observable là où on commence à prédire des valeurs non constamment nulles. Des pistes de solution seront évoquées dans la suite du rapport.

	$\hat{Y}_k^{(i)} > 0$	$\hat{Y}_k^{(i)} = 0$
$Y_k^{(i)} > 0$	Vrai Positif (VP)	Faux Négatif (FN)
$Y_k^{(i)} = 0$	Faux Positif (FP)	Vrai Négatif (VN)

TABLE 2.2 – Description des différentes catégories de classification.

	$\hat{Y}_k^{(i)} > 0$	$\hat{Y}_k^{(i)} = 0$
$Y_k^{(i)} > 0$	4740	639
$Y_k^{(i)} = 0$	11936	36985

TABLE 2.3 – Matrice de confusion pour la régression linéaire simple.

Ensemble	Médiane	Moyenne	Maximum
Entraînement	0.026	0.76	2.93
Test	0.033	0.77	2.98

TABLE 2.4 – Répartition de l'erreur J_k^{log} pour la prédiction de la moyenne. Les 150 espèces sont prises en compte. En moyenne, il y a un rapport $10^{0.77} \simeq 5.9$ entre les données et l'estimation.

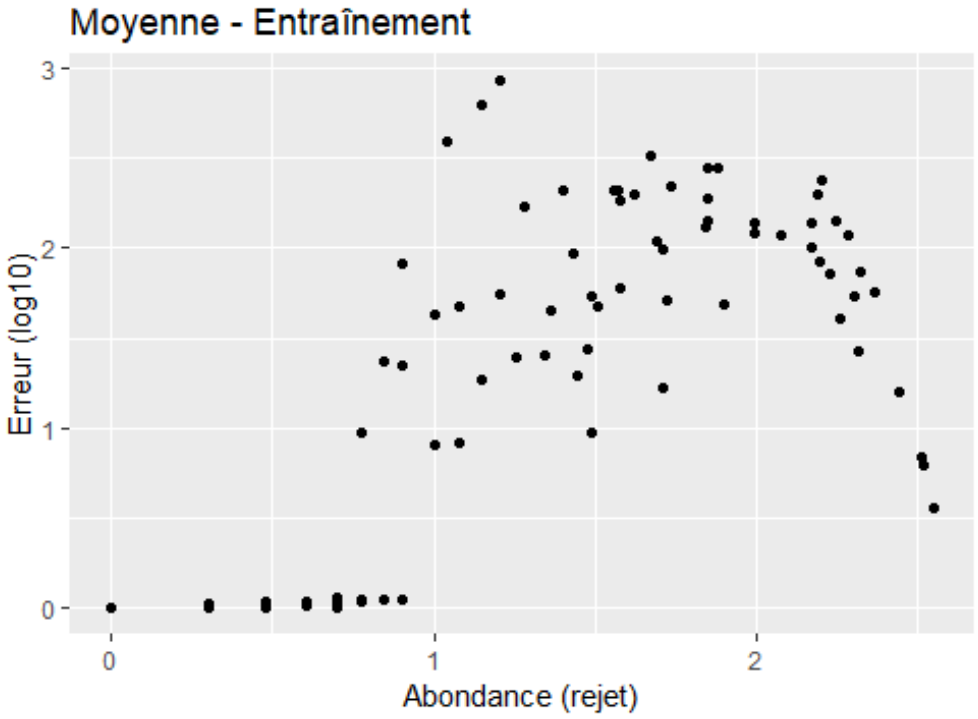

FIGURE 2.3 – Erreur J_k^{log} pour la prédiction de la moyenne

Transformation des données

La première solution pour prendre en compte les résidus croissants avec la quantité est de lisser les données en considérant leur logarithme. En effet, on constate que les ordres de grandeur des quantités observées varient beaucoup (de 100 à 10^5 par exemple). Il est possible que la régression fonctionne mieux pour des données plus homogènes. Cela peut être fait sur l'ensemble des données ou seulement sur les variables à expliquer : dans ce cas, on posera pour $k \in \{1, ..., K\}$: $\tilde{Y}_k = \log_{10}(Y_k + 1)$, pour éviter de prendre le logarithme d'une valeur nulle.

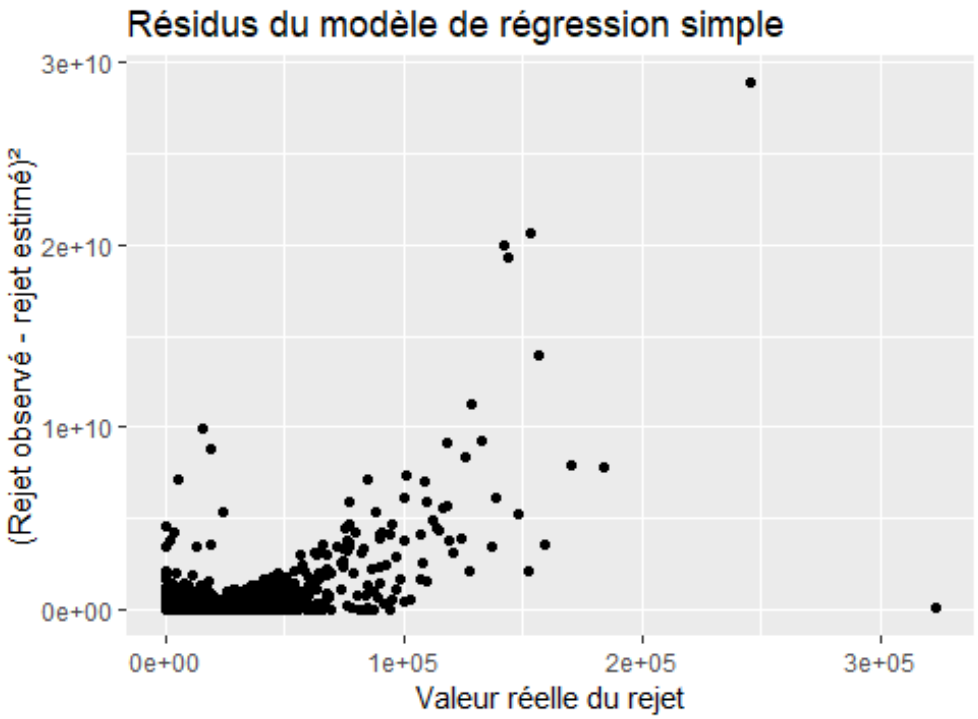

FIGURE 2.4 – Résidus du modèle linéaire simple en fonction de la quantité à prédire. On observe une dépendance positive : l'erreur est plus importante quand le rejet est plus conséquent. un point est une observation d'une espèce.

5 Une piste d'amélioration : le modèle joint

5.1 Présentation

Une autre idée est développée dans cette partie. Une explication simple à l'erreur élevée du modèle linéaire peut être proposée : en prédisant les quantités rejetées de manière indépendante pour chaque espèce, on se prive de l'information que nous apporte la corrélation entre les rejets eux-mêmes. Cette approche peut être mise en relation avec des modèles graphiques, qui estiment la matrice de covariance des variables. En particulier, la procédure Meinshausen-Bühlmann repose sur la sélection de voisinage variable par variable avec le LASSO (Meinshausen & Bühlmann, 2006). Elle est aussi réminiscente du *graphical-lasso* (Friedman et al., 2008), dont elle constitue en fait une forme simplifiée. On souhaite donc modifier le modèle linéaire en écrivant pour chaque espèce :

$$Y_k^{(i)} = \sum_{l \neq k} \alpha_{k,l} Y_l^{(i)} + \sum_{l=1}^{K} \beta_{k,l} X_l^{(i)} + \beta_k^0 + \epsilon_k^{(i)}.$$

L'estimation des paramètres de régression ne pose pas de difficultés particulières par rapport au cas précédent. Cependant, le couplage des variables $(Y_k^{(i)})_k$ nous empêche d'accéder directement à leurs valeurs prédites. Si on pose

$$D = \begin{pmatrix} 1 & -\alpha_{1,2} & \dots & -\alpha_{1,n} \\ -\alpha_{2,1} & 1 & \dots & -\alpha_{2,n} \\ \vdots & \dots & \ddots & \dots \\ \dots & \dots & -\alpha_{n-1}^n & 1 \end{pmatrix},$$

on cherche alors l'estimateur $\hat{\mathbf{Y}}$ de $\mathbf{Y}$ sous la forme d'une solution du système linéaire suivant, englobant les n opérations de pêche et les K espèces rejetées :

$$D.{}^t\hat{\mathbf{Y}} = \beta.{}^t\mathbf{X} + \beta^0.\mathbb{1}.$$

Il faut que la matrice D soit inversible, ce qui revient à supprimer du modèle les espèces qui rendent la matrice singulière. Dans les faits, la matrice est inversible lorsque l'on prend en compte les 62 espèces avec assez de rejets dans l'ensemble d'entraînement.

5.2 Résultats

Les résultats sont encore moins concluants à ce stade de l'analyse. Comme on l'observe sur la figure 2.5 et dans la table 2.5, l'erreur est supérieure à celle obtenue par le modèle linéaire simple alors que ce nouveau modèle est bien plus complexe. Une fois encore, la méthode **step** est utilisée pour choisir le meilleur groupe de variables selon l'AIC.

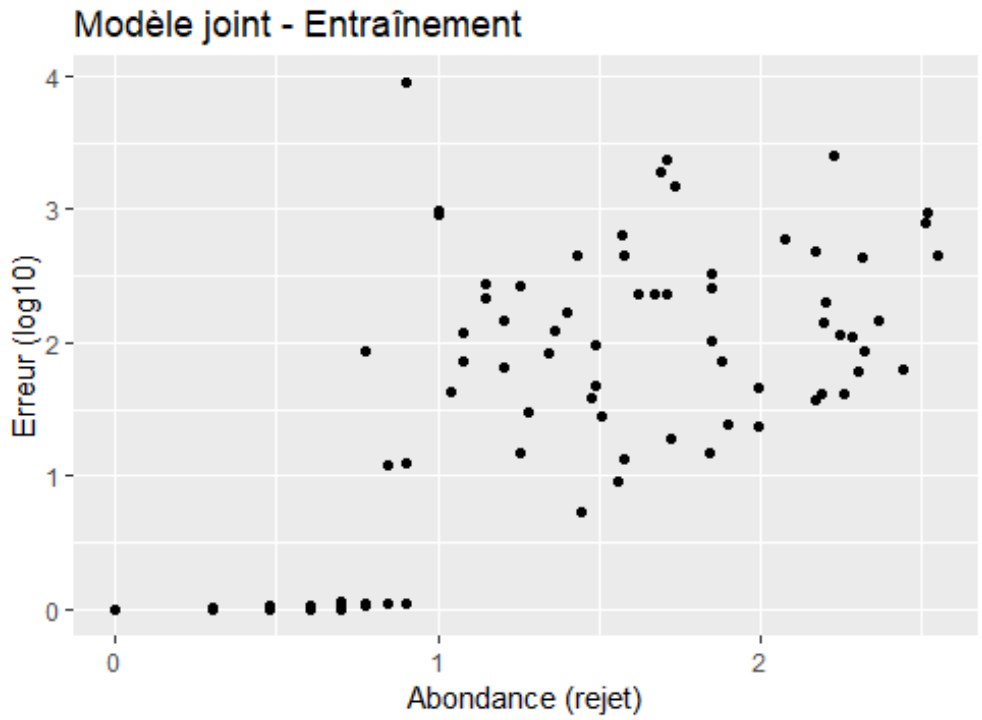

FIGURE 2.5 – Erreur J_k^{log} pour le modèle de régression linéaire joint avec sélection de variables AIC. La répartition est plus uniforme selon l'abondance.

Ensemble	Médiane	Moyenne	Maximum
Entraînement	0.026	0.88	3.95
Test	0.033	0.88	4.28

TABLE 2.5 – Répartition de l'erreur J_k^{log} pour le modèle de régression linéaire simple avec sélection de variables AIC. Les 150 espèces sont prises en compte. En moyenne, il y a un rapport $10^{0.88} \simeq 7.6$ entre les données et l'estimation (dans un sens ou dans l'autre).

Une fois encore, on peut calculer la fraction de coefficients de régression conservés par le modèle : on retient 715 coefficients sur les $62 \times (61 + 90)$ possibles, soit 7.6% pour les variables de rejets et

explicatives. Enfin, on peut évaluer la prédiction de présence. La prédiction est juste dans 79% des cas, avec une sensibilité de seulement 43% et une spécificité de 82%.

Comparaison à la régression linéaire simple Comme la matrice D est inversible, on peut réécrire le problème sous la forme :

$$^t\hat{\mathbf{Y}} = D^{-1}(\beta.{}^t\mathbf{X} + \beta^0.\mathbb{1}).$$

Or le modèle de régression linéaire simple s'écrit (avec d'autres paramètres de régression $\beta', (\beta^0)'$) :

$$^t\hat{\mathbf{Y}} = \beta'.{}^t\mathbf{X} + \beta'^0.\mathbb{1}.$$

Ainsi, on résout un problème identique, avec $D^{-1}\beta = \beta'$ et $D^{-1}\beta^0 = \beta'^0$! En fait, le modèle couplé prend en compte plus de paramètres et devrait être mieux ajusté. Cependant, le modèle de prédiction nécessite la résolution d'un système linéaire. Cela implique des imprécisions qui ont un coût élevé sur les résultats finaux.

Conclusion partielle Il semble qu'on atteigne ainsi les limites intrinsèques au modèle linéaire : la structure des données ne convient pas à ses hypothèses, notamment concernant la distribution des erreurs. Nous proposons donc de compléter cette modélisation avec une prédiction préalable de présence ou d'absence.

Chapitre 3

Modèle à deux étapes

1 Principe

1.1 Modèle à obstacle

Le *modèle à deux étapes* que l'on a mis au point est basé sur un modèle à obstacle ou *hurdle model*. Dans ce type de modèle, il y a une certaine probabilité d'observer des valeurs nulles : le processus sous-jacent doit d'abord franchir un "obstacle" (Cragg, 1971). Cette idée s'applique dans beaucoup de domaines : le modèle à obstacle est largement utilisé dans les différentes disciplines de la biologie (Heilbron, 1994 ; McDavid et al., 2019) et de l'écologie (Welsh et al., 1996 ; Martin et al., 2005 ; Cunningham & Lindenmayer, 2005 ; Wenger & Freeman, 2008).

Ici, ce modèle permet de prendre en compte l'excès de zéros. La probabilité d'observer une valeur non-nulle est la probabilité que l'obstacle soit franchi et dans ce cas, la quantité suit une distribution connue, par exemple une loi de Poisson ou une loi gaussienne. Ainsi, le modèle s'obtient en ajoutant une masse de Dirac en 0 à une loi de probabilité $\mathcal{G}$. Si g est la densité de $\mathcal{G}$ par rapport à la mesure de Lebesgue et $p \in]0,1[$ la probabilité de franchir l'obstacle, alors la loi du modèle à obstacle est définie par :

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

f est bien une densité par rapport à λ_0, somme de la mesure de Lebesgue et de la mesure de Dirac en 0.

1.2 Décomposition de la vraisemblance

Pour alléger les notations, on considère une seule espèce de poisson rejetée k et on note $Y := Y_k$. Si Y suit un modèle à obstacle de densité f, alors on introduit la variable aléatoire binaire observée V définie par $V = \mathbb{1}_{Y \neq 0}$. On en déduit que la probabilité de franchir l'obstacle est $p = \mathbb{P}(V = 1)$ avec :

$$V \sim \mathcal{B}(p),$$

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

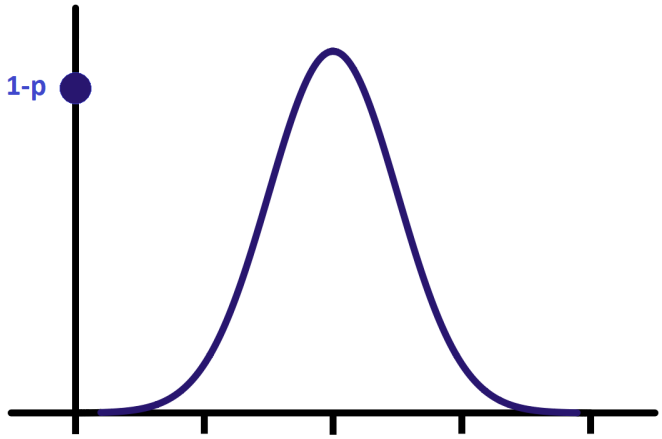

FIGURE 3.1 – Distribution d'un modèle à obstacle issu d'une loi normale.

Pour une observation $(Y^{(i)}, V^{(i)})$, la vraisemblance s'écrit :

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

On en déduit la log-vraisemblance :

$$\mathcal{LL}(Y^{(i)}, V^{(i)}) = \log\{\mathbb{1}_{V^{(i)}=0}(1-p) + \mathbb{1}_{V^{(i)}=1}p \times g(Y^{(i)})\}.$$

La log-vraisemblance complète pour l'ensemble des observations s'écrit :

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

Enfin, si on estime $V^{(i)}$ et $Y^{(i)}$ à partir de $X^{(i)}$ avec des paramètres indépendants γ et β, alors en notant $p(X^{(i)};\gamma) := \mathbb{P}(V^{(i)}=1|X^{(i)};\gamma)$, on a :

$$\begin{aligned}\mathcal{LL}(\mathbf{Y}, \mathbf{V}|\mathbf{X};\beta,\gamma) &= \sum_{i:V^{(i)}=0} \log(1-p(X^{(i)};\gamma)) + \sum_{i:V^{(i)}=1} \log p(X^{(i)};\gamma)\\ &\quad + \sum_{i:Y^{(i)}\neq 0} \log g(Y^{(i)}|X^{(i)};\beta)\\ &= \sum_{i=1}^{n} \left\{(1-V^{(i)})\log(1-p(X^{(i)};\gamma)) + V^{(i)}\log p(X^{(i)};\gamma)\right\}\\ &\quad + \sum_{i\in J} \log g(Y^{(i)}|X^{(i)};\beta)\\ &= \mathcal{LL}_1(\mathbf{V}|\mathbf{X};\gamma) + \mathcal{LL}_2(\mathbf{Y}^{(J)}|\mathbf{X}^{(J)};\beta),\end{aligned}$$

avec $\mathbf{Y}^{(J)} = (Y^{(i)})_{i\in J}$, $\mathbf{X}^{(J)} = (X^{(i)})_{i\in J}$ et $J = \text{supp}(\mathbf{Y})$.

On peut donc maximiser la vraisemblance en traitant séparément $\mathcal{LL}_1$ et $\mathcal{LL}_2$, car elles dépendent de paramètres indépendants. Pour $\mathcal{LL}_1$, on reconnaît la forme de la log-vraisemblance pour un échantillon de variables aléatoires indépendantes et distribuées suivant des lois de Bernoulli de paramètres respectifs $p^{(i)} = p(X^{(i)};\gamma)$. On cherche $\hat{\gamma}$ qui la maximise. Pour $\mathcal{LL}_2$, on reconnaît la log-vraisemblance pour les variables $Y^{(i)}$ telles que $Y^{(i)} \neq 0$. On détermine alors $\hat{\beta}$ en ne prenant que les données non-nulles.

Grâce au modèle à obstacle, on a décomposé le problème en deux étapes découplées. Ainsi, on définit le *modèle à deux étapes*, un modèle d'inférence, à partir de la décomposition du modèle à obstacle. On va combiner :

— l'étape C (classification), qui réalise une classification binaire pour prédire la probabilité que la donnée soit non-nulle,

— l'étape R (régression), qui effectue une régression pour les données qu'on a prédites non-nulles.

Entre les deux étapes, on met en place une stratégie décisionnelle, à savoir des seuils sur les probabilités sorties de l'étape C pour décider si elles correspondent à des données nulles. Il faut entraîner un modèle pour chacune des deux étapes, puis choisir une stratégie pour fixer les seuils. Dans le paragraphe suivant, on construit un modèle basique pour les deux étapes.

1.3 Exemple détaillé

On ajuste un modèle de régression logistique pour l'étape C et un modèle de régression linéaire pour l'étape R. Dans un premier temps, on ne considère pas de régularisation.

Étape C : Régression logistique

On veut prédire la présence ou non de l'espèce dans les rejets. Si on note $p^{(i)} := \mathbb{P}(V^{(i)} = 1|X^{(i)})$, alors le problème de régression logistique peut s'écrire :

$$\log \frac{p^{(i)}}{1-p^{(i)}} = \gamma^T x^{(i)},$$
$$p^{(i)} = \frac{\exp(\gamma^T x^{(i)})}{1+\exp(\gamma^T x^{(i)})}.$$

Puis, l'estimateur $\hat{\gamma}$ est obtenu en maximisant la log-vraisemblance :

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmax}}\, \mathcal{LL}_1(\gamma),$$

avec :

$$\begin{aligned}\mathcal{LL}_1(\gamma) &= \sum_{i=1}^{n} \left\{ v^{(i)} \log p^{(i)} + (1-v^{(i)}) \log(1-p^{(i)}) \right\} \\ &= \sum_{i=1}^{n} \left\{ v^{(i)} \gamma^T x^{(i)} - \log(1+\exp(\gamma^T x^{(i)})) \right\}.\end{aligned}$$

Ainsi, avec $\hat{\gamma}$ et un vecteur de variables $x^{(i)}$, on peut estimer la probabilité que $V^{(i)}$ soit égal à 1 :

$$\hat{p}^{(i)} = \frac{\exp(\hat{\gamma}^T x^{(i)})}{1+\exp(\hat{\gamma}^T x^{(i)})}.$$

Finalement, on peut se fixer un certain seuil de décision μ pour estimer $\hat{V}^{(i)} = \mathbb{1}_{\hat{p}^{(i)}>\mu}$. Naturellement, le seuil qui minimise la probabilité de se tromper est $\mu = 1/2$. Dans la suite, on discutera des cas où il est intéressant de choisir une autre valeur.

Étape R : Régression linéaire

Pour la régression linéaire, on utilise comme d'habitude la méthode du maximum de vraisemblance. Soit $J = \text{supp}(\mathbf{Y})$:

$$\mathbf{Y}^{(J)} = \mathbf{X}^{(J)}\beta + \epsilon,$$

avec $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{|J|})$, alors la log-vraisemblance s'écrit :

$$\mathcal{LL}_2(\beta) = \sum_{i \in J} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y^{(i)} - \beta^T x^{(i)})^2}{2\sigma^2} \right\}.$$

En choisissant l'estimateur du maximum de vraisemblance :

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\, \mathcal{LL}_2(\beta),$$

on obtient un estimateur de $Y^{(i)}$:

$$\hat{Y}^{(i)} = \hat{\beta}^T x^{(i)}.$$

Cet estimateur a été seulement ajusté à partir des données $\mathbf{Y}^{(J)}$, c'est-à-dire telles que $Y^{(i)}$ soit non-nul. Il ne devrait donc être utilisé que pour les données où l'on a décidé à l'étape R que $\hat{V} = 1$.

Ainsi, dans le cas général avec $\tilde{K} = 120$ espèces, il faut entraîner un modèle d'inférence pour chaque espèce et chaque étape. Dans cet exemple, on a choisi deux modèles simples pour les deux étapes, mais on pourrait les remplacer par d'autres modèles statistiques judicieux. De futures études pourraient être conduites sur d'autres choix de modèle.

2 Stratégies d'application

Une fois les modèles des étapes C et R ajustés, on peut enfin les combiner pour appliquer le modèle à deux étapes aux données d'ObsMer. Dans cette section, on présente plusieurs stratégies d'application. Le choix d'une stratégie correspond au choix des méthodes de classification pour l'étape C, de régression pour l'étape R et de choix d'un seuil de décision μ.

2.1 Choix du seuil de décision

Dans ce paragraphe, on présente les principales méthodes pour le choix des seuils de décision. On va d'abord donner des arguments dans le cas d'une seule espèce k, puis on décrira deux méthodes de généralisation pour le cas de plusieurs espèces.

Classificateur *plug-in*

On considère encore une seule espèce k et $Y := Y_k$. Pour la i-ème observation $X^{(i)}$, on écrit $p^{(i)}(x) = \mathbb{P}(V^{(i)} = 1 | X^{(i)} = x)$, en cohérence avec les notation précédentes. Le classifieur de Bayes pour l'estimation de $V^{(i)}$ est défini par :

$$f^*_{V^{(i)}}(x) = \begin{cases} 1 & \text{si } p^{(i)}(x) \geq 1/2, \\ 0 & \text{sinon .} \end{cases}$$

$f^*_{V^{(i)}}$ est le classifieur qui minimise la probabilité de se tromper (Devroye et al., 2013). Ainsi, pour minimiser le taux d'erreur, il faudrait choisir $\hat{V}^{(i)} = f^*_{V^{(i)}}(x^{(i)})$.

Comme on l'a vu précédemment, les méthodes de classification probabilistes telle que la régression logistique estiment $p^{(i)}$. Il est alors possible de remplacer $p^{(i)}$ par $\hat{p}^{(i)}$ dans le classifieur de Bayes pour obtenir un classifieur *plug-in*, s'écrivant sous la forme :

$$f_{V^{(i)}}(x) = \begin{cases} 1 & \text{si } \hat{p}^{(i)}(x) \geq 1/2, \\ 0 & \text{sinon .} \end{cases}.$$

Cela revient à choisir un seuil $\mu = 1/2$.

Le classificateur *plug-in* est un choix naturel, mais il nécessite que l'estimation de $\hat{p}$ soit sans biais. C'est bien le cas dans la régression logistique qu'on a décrite précédemment. En revanche, dans la suite on va introduire une régularisation l_1 qui amène à un biais. Il faudra donc faire attention au choix du seuil.

Métriques de classification

Une autre méthode pour choisir le seuil de décision est de regarder les mesures de la performance de la classification. Dans notre cas, la classification binaire porte sur la prédiction de $V \in \{0, 1\}$. Une prédiction correcte peut correspondre à deux issues, un vrai positif (VP) ou un vrai négatif (VN) et une erreur à deux autres, un faux positif (FP) ou un faux négatif (FN). On peut comparer les fréquences de ces issues, par exemple avec la sensibilité et la spécificité.

Comme l'objectif est de minimiser FN et FP, on veut avoir la sensibilité et la spécificité les plus proches de 1 possible. Le choix du seuil de décision change les prédictions de $\hat{V}$ et influe donc le nombre de faux négatifs et de faux positifs. Si le seuil μ augmente, on prédit plus de négatifs, donc VN et FN augmentent et VP et FP diminuent. Ainsi, la spécificité augmente et la sensibilité diminue. On se retrouve alors face à un dilemme : augmenter l'une des métriques en faisant varier μ réduit mécaniquement l'autre. Il faut donc définir une fonction des deux quantités FN et FP à optimiser.

Fonction de coût

Un choix raisonnable est de minimiser $J(\mu) = \mathrm{FP}(\mu) + \mathrm{FN}(\mu)$ pour minimiser le taux d'erreur total. Cependant, on veut parfois éviter à tout prix d'avoir des faux positifs ou des faux négatifs, quitte à admettre un plus grand taux d'erreur total. Dans ce cas, il est judicieux d'attribuer des poids (ω_{FP} et ω_{FN}) à chaque issue pour pondérer leur importance. Ensuite, on minimise $J(\mu) = \omega_{\mathrm{FP}}.\mathrm{FP}(\mu) + \omega_{\mathrm{FN}}.\mathrm{FN}(\mu)$. Par exemple, lors du test pour une maladie grave, un choix cohérent serait $\omega_{\mathrm{FP}} = 1$ et $\omega_{\mathrm{FN}} = 1000$ (Bishop, 2006) : on veut être sûr de détecter tous les malades (FN signifiant ici malade mais dont le test est négatif).

Dans notre problème, le but ultime est de prédire des quantités pour les rejets. La classification n'est qu'une étape intermédiaire du modèle qui est censée nous faciliter la tâche en palliant le problème de l'excès de zéros. L'influence des faux positifs et des faux négatifs n'est pas évidente et leur coût peut varier en fonction de l'espèce. Certains faux négatifs peuvent être préférables à des vrais positifs : pour des observations où Y est faible mais non-nul, il se peut que le modèle de régression de l'étape R ne puisse pas prédire de valeur assez petite. Dans ce cas, il vaut mieux prédire 0 et de payer le coût de la classification, plutôt que de passer par l'étape R et de payer le coût de la régression.

On définit donc le coût final du modèle à deux étapes à partir des erreurs de prédiction finale. Soit f^{μ} le régresseur du modèle à deux étapes avec le seuil de décision μ : pour une certaine réalisation

$x^{(i)}$ de $X^{(i)}$, on prédit $\hat{Y}^{(i)} = f^{\mu}(x^{(i)})$. Pour une métrique Δ, le coût associé peut alors s'écrire :

$$\begin{aligned} J(\mu) &= \sum_{i=1}^{n} \Delta(y^{(i)}, f^{\mu}(x^{(i)})) \\ &= \sum_{i:f^{\mu}(x^{(i)})=0} \Delta(y^{(i)}) + \sum_{i:y^{(i)}=0} \Delta(f^{\mu}(x^{(i)})) + \sum_{\substack{i:y^{(i)}\neq 0,\\ f^{\mu}(x^{(i)})\neq 0}} \Delta(y^{(i)}, f^{\mu}(x^{(i)})) \\ &= J_{FN}(\mu) + J_{FP}(\mu) + J_{TP}(\mu). \end{aligned}$$

Le coût peut donc être décomposé :
— pour les vrais négatifs, on ne paie pas de coût,
— pour les faux négatifs, on prédit zéro à tort, donc le coût dépend de la valeur de Y,
— pour les faux positifs, l'étape R prédit une valeur pour Y alors que celle-ci est nulle, donc le coût dépend de la valeur prédite $\hat{Y} = f^{\mu}(x)$,
— pour les vrais positifs, le coût est l'erreur de la régression de l'étape R.

Ainsi, il n'est pas nécessaire d'attribuer des poids arbitraires aux différentes issues. Elles sont automatiquement pondérées dans J_{FN}, J_{FP} et J_{TP}. Cela confère un critère objectif pour choisir les seuils de décision.

Méthode jointe, méthode séparée

On revient maintenant au cas général avec K espèces ou plutôt à $\tilde{K} = 120$ espèces en retirant les espèces jamais rejetées (chapitre 2, section 1). On a donc $\tilde{K}$ modèles de classification pour l'étape C, $\tilde{K}$ modèles de régression pour l'étape R, $\tilde{K}$ seuils de décisions et donc $\tilde{K}$ régresseurs f_k et $\tilde{K}$ fonctions de coût J_k.

On a deux possibilités pour le choix des seuils μ_k. On peut imposer de prendre le même pour toutes les espèces (méthode jointe). Dans ce cas, on a $\mu_k = \tilde{\mu}$ pour $k = 1, ..., \tilde{K}$ et on optimise une seule fonction :

$$\begin{aligned} J(\tilde{\mu}) &= \sum_{k=1}^{\tilde{K}} J_k(\tilde{\mu}) \\ &= \sum_{k=1}^{\tilde{K}} \sum_{i=1}^{n} \Delta(y_k^{(i)}, f_k^{\tilde{\mu}}(x^{(i)})). \end{aligned}$$

Sinon, on autorise les espèces à avoir des seuils différents (méthode séparée) et pour chaque $k = 1, ..., \tilde{K}$, on optimise :

$$J_k(\mu_k) = \sum_{i=1}^{n} \Delta(y_k^{(i)}, f_k^{\mu_k}(x^{(i)})).$$

Les résultats de ces deux approches seront comparés dans une section ultérieure.

2.2 Choix des modèles pour les étapes C et R

L'exemple détaillé (section 1.3) donne déjà deux modèles simples pour ces étapes, la régression logistique pour l'étape C et la régression linéaire pour l'étape R. Ces modèles utilisent toutes les variables explicatives ($p = 154$) et sont donc trop complexes. On procède à une sélection de variables par pénalisation.

La pénalisation l_1 a déjà été présentée dans le cadre d'un modèle linéaire (chapitre 2, section 2). On peut utiliser la même méthode pour la régression logistique maximisant la log-vraisemblance

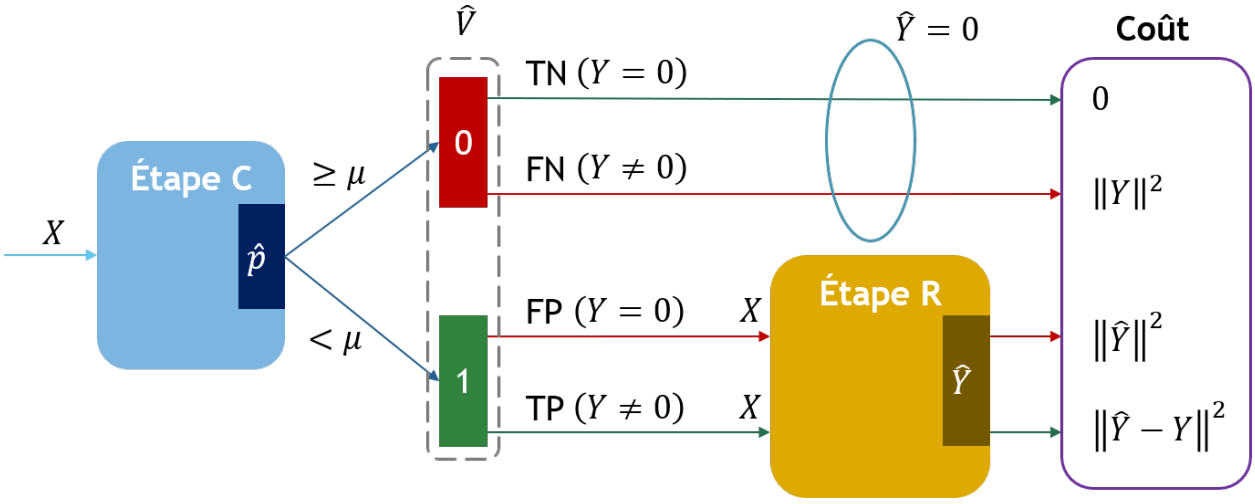

FIGURE 3.2 – Schéma récapitulatif de la procédure d'inférence du modèle à deux étapes, avec les différents types de coût.

pénalisée :

$$\mathcal{LL}'(\gamma) = \sum_{i=1}^{n} \left\{ v^{(i)} \gamma^T x^{(i)} - \log(1 + \exp(\gamma^T x^{(i)})) \right\} - \lambda \sum_{j=1}^{p} |\gamma_j|.$$

Plus le coefficient de régularisation λ est grand, plus on force la parcimonie dans $\hat{\gamma}$. La valeur de λ qui donne la vraisemblance la plus élevée est $\lambda = 0$. Il faut donc choisir sa valeur en regardant la performance du modèle sur des données nouvelles et non sur celles d'entraînement.

Pour cela, on va faire appel à la validation croisée V-fold. On découpe l'ensemble des données d'entraînement en V parties. Pour chaque partie, on teste les modèles que l'on a ajustés sur le reste des données, ce qui donne V courbes d'erreur en fonction de λ. Un premier critère, celui du minimum, est de choisir $\lambda = \lambda_{min}$ qui minimise la moyenne de ces fonctions d'erreur.

On peut aussi appliquer le critère du *one-standard-error* (Hastie et al., 2009). Dans ce cas, on considère que l'estimation de la fonction d'erreur est imprécise et on choisit dans le doute un λ plus grand pour induire plus de parcimonie dans γ. On choisit alors la plus grande valeur de $\lambda = \lambda_{1se}$ qui reste à un écart-type (figure 3.3) du minimum. On comparera plus tard ce critère au critère du minimum.

Choix	Modèle C	Seuil de décision	Modèle R
Stratégie	λ^C_{min}	Jointe	λ^R_{min}
	λ^C_{1se}	Séparée	λ^R_{1se}

TABLE 3.1 – Récapitulatif des différentes stratégies à possibles. Il y a deux possibilités pour chaque étape du modèle, mais également pour le choix du seuil de décision.

3 Application

Le modèle à deux étapes est appliqué aux données d'ObsMer avec toutes les variantes décrites dans la section précédente, qui seront comparées une à une, c'est-à-dire pour la régression logistique de l'étape C et la régression linéaire de l'étape R, les deux critères de sélection de λ le coefficient de régularisation l_1, ainsi que les deux méthodes de sélection du seuil de décision (table 3.1).

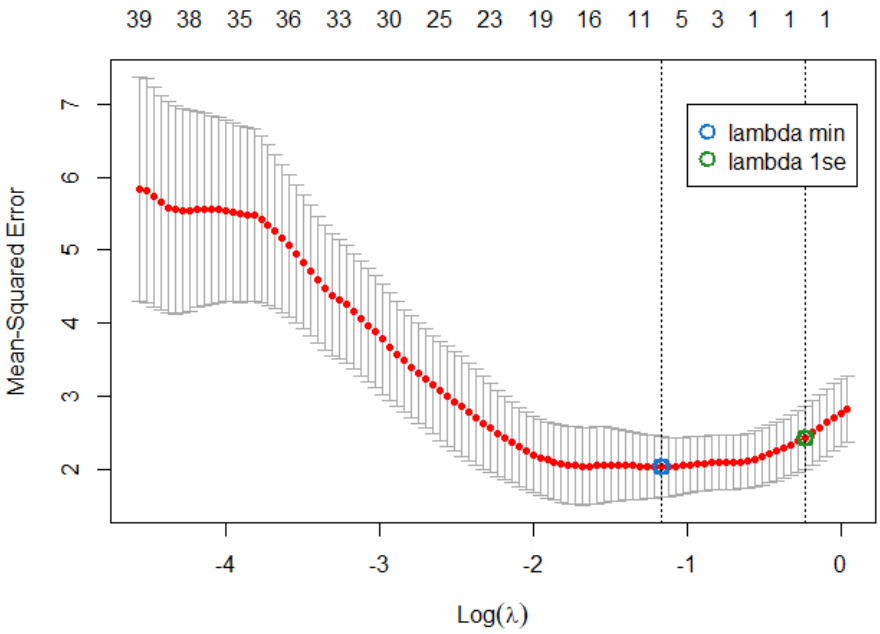

FIGURE 3.3 – Coût de validation croisée en fonction du coefficient λ pour le modèle de l'étape R de Y_{60}, les rejets de la cardine franche. La sélection de λ avec les critères du minimum et du *one-standard-error* est automatique.

3.1 Implémentation

Le processus complet d'inférence avec le modèle à deux étapes comporte plusieurs parties :

1. Transformation logarithmique des quantités débarquées $\tilde{Y}_k \leftarrow \log_{10}(Y_k + 1)$, pour $k = 1, ..., K$;
2. Création d'un ensemble d'entraînement et d'un ensemble de test ;
3. Ajustement du modèle de classification de l'étape C avec toutes les données de l'ensemble d'entraînement pour chaque espèce k ;
4. Ajustement du modèle de régression de l'étape R avec les données non-nulles de l'ensemble d'entraînement pour chaque espèce k ;
5. Pour chaque élément d'une suite de seuils $(\mu_h)_{1 \leq h \leq H}$:
   (a) Application de l'étape C à l'ensemble de test pour trouver ses probabilités $\hat{\mathbf{p}}_k$ pour chaque espèce k ;
   (b) Application du seuil de décision sur $\hat{\mathbf{p}}_k$ pour avoir $\hat{\mathbf{V}}_k$ pour chaque espèce k ;
   (c) Pour chaque espèce k, pour chaque observation i de l'ensemble de test :
   — si $\hat{V}_k^{(i)} = 0$, alors on prédit $\hat{\hat{Y}}_k^{(i)} = 0$;
   — si $\hat{V}_k^{(i)} = 1$, alors on applique l'étape R pour calculer $\hat{\hat{Y}}_k^{(i)}$;
   (d) Calcul des erreurs de prédiction ;
6. Sélection des seuils μ_k, pour chaque k en fonction des erreurs de prédiction, selon la méthode choisie (jointe ou séparée).

Le package **R** `glmnet` déjà mentionné précédemment, propose des solutions pour les méthodes de vraisemblance pénalisée par la norme l_1, à la fois pour la régression logistique et la régression

linéaire. Il propose aussi de prendre en charge l'étape de validation croisée et de sélectionner de manière automatique les coefficients λ avec les critères du minimum et du *one-standard-error*. La procédure V-fold requiert notamment un minimum de V observations non-nulles. On a choisi $V = 5$, donc , comme évoqué plus haut, on entraîne les modèles de C et R pour les espèces pour lesquelles il y a au moins 5 rejets non-nuls dans l'ensemble d'entraînement et on prédira toujours des valeurs nulles pour les espèces qui ne franchissent pas ce seuil.

L'erreur de prédiction est calculée avec la distance logarithmique $\Delta(\hat{Y}, Y) = |\log_{10} \frac{\hat{Y}+1}{Y+1}|$ (chapitre 2, section 4.1). Avec les données transformées, son expression se simplifie : $\Delta(\hat{Y}, Y) = |\hat{\tilde{Y}} - \tilde{Y}|$. On définit l'erreur moyenne logarithmique par espèce :

$$J_k^{\log} = \frac{1}{n} \sum_{i=1}^{n} \Delta(\hat{Y}_k^{(i)}, Y_k^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} |\hat{\tilde{Y}}_k^{(i)} - \tilde{Y}_k^{(i)}|.$$

L'erreur logarithmique moyenne de l'ensemble s'écrit :

$$\begin{aligned} J^{\log} &= \tfrac{1}{K} \sum_{k=1}^{K} J_k^{\log} \\ &= \tfrac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} \Delta(\hat{Y}_k^{(i)}, Y_k^{(i)}) \\ &= \tfrac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} |\hat{\tilde{Y}}_k^{(i)} - \tilde{Y}_k^{(i)}|. \end{aligned}$$

Comme on l'a vu précédemment, cette quantité peut s'interpréter comme une moyenne géométrique de l'erreur en d'ordre de grandeur. Cette erreur est calculée pour une suite de seuils μ_h. Finalement, on utilise la méthode séparée ou la méthode jointe présentées précédemment pour fixer les seuils de chaque espèce.

3.2 Résultats

Pour la méthode jointe de sélection du seuil, on a tracé J^{log} pour différentes valeurs du seuil, que l'on compare aux valeurs obtenues avec la méthode séparée et à plusieurs références. Les premières références sont obtenues en calculant la moyenne des rejets $\bar{\tilde{y}}_k^{train}$ pour chaque espèce dans l'ensemble d'entraînement. La prédiction se fait alors avec ces moyennes, indépendamment des variables explicatives, c'est-à-dire en utilisant le régresseur $f_{B1}(x^{(i)}) = \bar{\tilde{y}}^{train}$, pour l'observation i. On remarque que i n'apparaît pas dans le membre de droite, donc quelque soit l'observation, la valeur prédite est la même. En fait, elle ne dépend que de la partition des données initiales en ensembles d'entraînement et de test. Les erreurs logarithmiques moyennes de ce régresseur pour les ensembles d'entraînement et de test sont notées B_1^{train} et B_1^{test}.

Au lieu de calculer la moyenne sur toutes les données d'entraînement, on peut le faire pour les données non-nulles seulement ($\bar{\tilde{y}}_k^{train \neq 0}$). Une fois les modèles de l'étape C et R ajustés, on choisit les seuils par la méthode séparée. Puis on remplace l'étape R par la prédiction de ces moyennes. Les erreurs logarithmiques moyennes sont notées B_2^{train} et B_2^{test}. Contrairement à B_1^{train} et B_1^{test}, ces références dépendent des étapes C et R. C'est évident pour C, et elles dépendent aussi des seuils puisque l'étape R n'est appliquée que si $\hat{V}_k^{(i)} = 1$.

Ces valeurs de référence sont utiles pour évaluer la performance du modèle à deux étapes. Le régresseur f_{B1} est extrêmement simple. Si les erreurs sont plus grandes que B_1^{train} et B_1^{test}, le modèle est particulièrement mauvais. B_2^{train} et B_2^{test} évaluent principalement la performance de l'étape C. Si elles sont notablement inférieures à B_1^{train} et à B_1^{test}, alors la prédiction des zéros avec $\hat{V}_k^{(i)} = 0$ diminue l'erreur de manière significative.

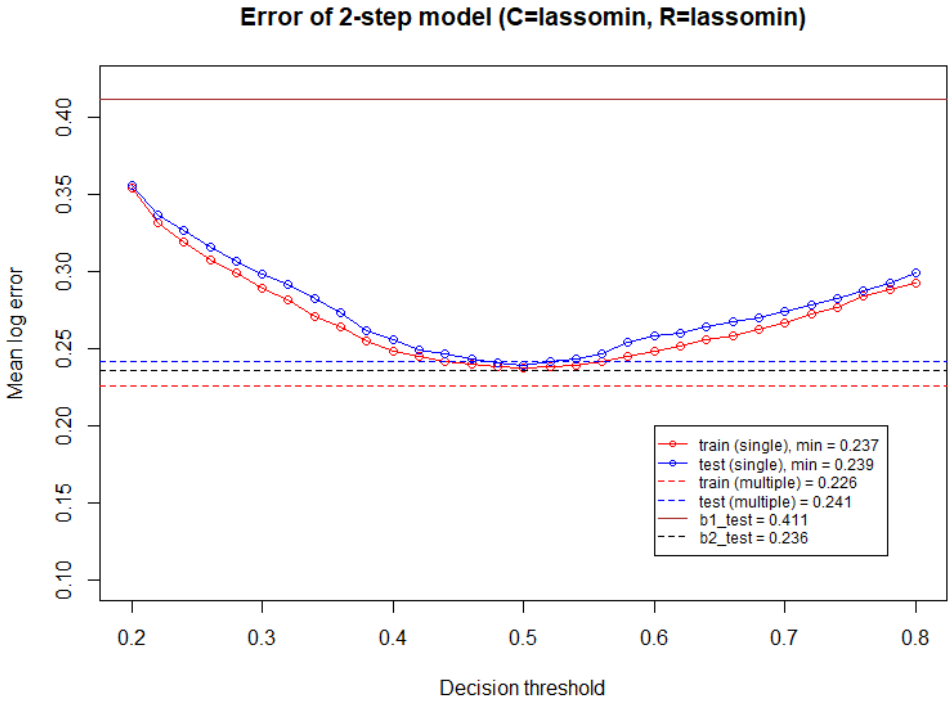

FIGURE 3.4 – Erreurs moyennes logarithmiques en choisissant les coefficients de régularisation avec le critère du minimum pour les étapes C et R et les seuils avec les méthodes jointe (single) et séparée (multiple).

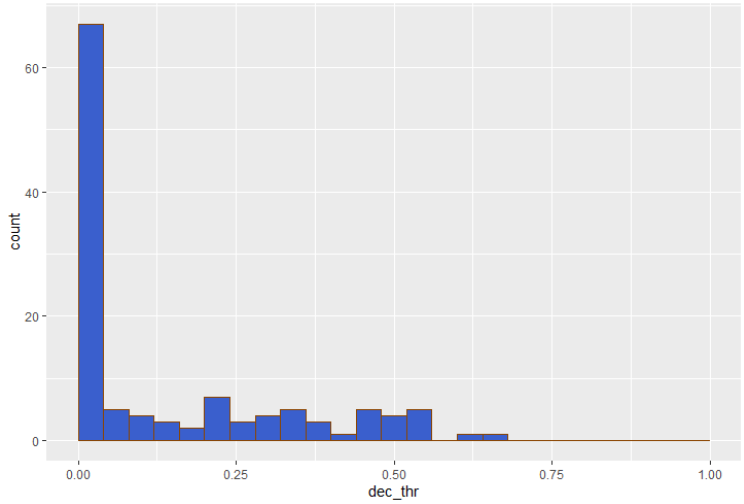

FIGURE 3.5 – Distribution des seuils de décision choisis pour les 150 espèces par la méthode séparée. Dans le cas où plusieurs seuils sont équivalents, on choisit le plus petit.

On a d'abord utilisé le critère du minimum pour la sélection des coefficients de régularisation λ pour les deux étapes (figure 3.4). Les erreurs pour les deux méthodes de sélection des seuils (jointe et séparée) sont bien inférieures à B_1^{test}. Les courbes d'entraînement et de test pour la sélection jointe des seuils atteignent leur minimum pour un seuil de $\tilde{\mu} = 0.5$. Cela suggère que l'estimation des $\hat{p}_k^{(i)}$ est convenable. Pour les erreurs obtenues avec la méthode séparée, on constate que cette méthode entraîne une erreur d'entraînement plus petite que celles de la méthode jointe. D'autre part, l'erreur de test est plus élevée que pour certains seuils de la méthode jointe, donc la méthode séparée

surajuste les seuils. On constate dans la figure 3.5 qu'un seuil de 0.5 n'est pas très fréquemment choisi, contrairement au cas de la méthode jointe.

Il est plus frappant de comparer ces valeurs à B_2^{test}. Même avec le seuil joint de 0.5, ni l'erreur d'entraînement, ni celle de test ne descend en deçà. Ainsi, l'étape R ne remplit pas sa fonction : le modèle choisi n'est pas plus efficace que la prédiction de la moyenne. Dans le cas de la régression LASSO, les méthodes d'optimisation devraient réduire les coefficients à zéro pour prédire la moyenne. C'est ce qui se passe en réalité pour plusieurs espèces (figure A.10). Si l'erreur du LASSO est plus élevée que B_2^{test}, c'est que l'on n'entraîne un modèle pour une espèce que si celle-ci est présente au moins 5 fois dans les données d'entraînement, et on prédit toujours 0 dans l'étape R pour les autres. Ce coût, que l'on remarque ici, n'est pas élevé et la plupart du temps ces prédictions sont correctes car ces espèces ne sont présentes que peu de fois.

Avec le critère du *one-standard-error*, on obtient des résultats différents. L'erreur B_2^{test} est beaucoup plus élevée (figure A.7) qu'avec le critère du minimum, donc l'étape C est considérablement moins performant. On propose donc d'utiliser le critère du minimum pour l'étape C et le critère du *one-standard-error* pour l'étape R. À première vue, cela donne des résultats surprenants (figure 3.6). On retrouve quelques points communs avec l'utilisation du critère du minimum pour l'étape R. Avec la méthode jointe, le seuil qui minimise l'erreur est toujours 0.5 et la méthode séparée semble toujours surajuster le modèle.

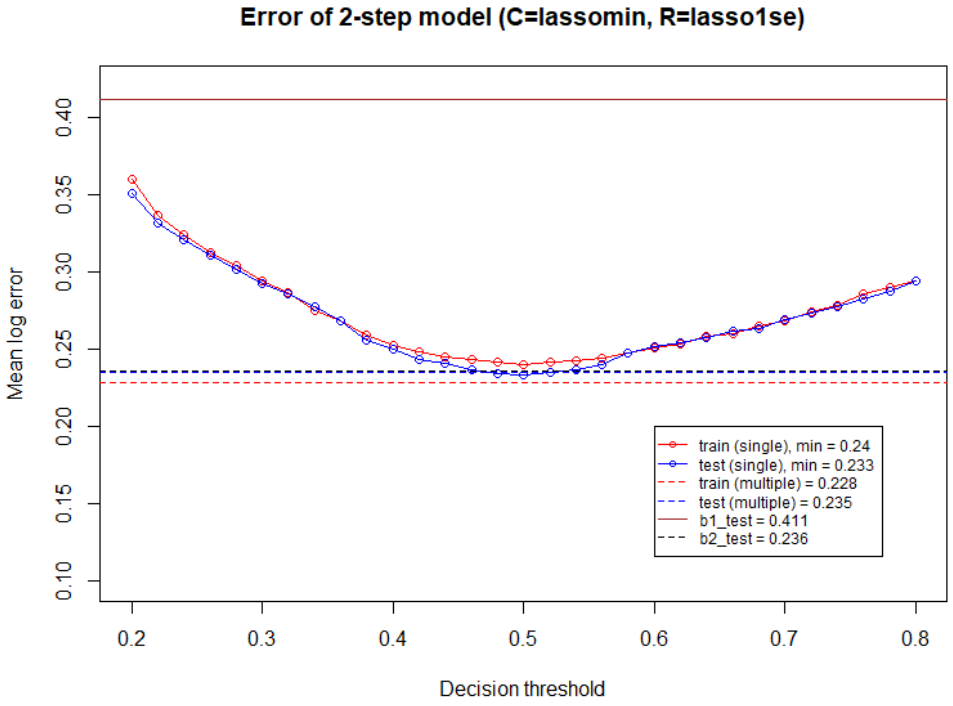

FIGURE 3.6 – Erreurs moyennes logarithmiques en choisissant les coefficients de régularisation avec le critère du minimum pour l'étape C et le critère du *one-standard-error* pour l'étape R, et les seuils avec les méthodes jointe (single) et séparée (multiple).

Le constat est similaire en analysant la distribution des erreurs des espèces (tables 3.2 et 3.3). On remarque néanmoins qu'il y a de plus grandes différences dans les maxima que dans les moyennes et les médianes qui sont des statistiques plus robustes, en particulier ici aux différences inter-espèces.

Par contre, les erreurs de test sont plus petites que précédemment. En particulier, elles sont plus basses que celles d'entraînement, de manière inattendue. On interprète cela en notant que

	Médiane	Moyenne	Maximum
Train	0.049	0.228	1.260
Test	0.076	0.234	1.252

TABLE 3.2 – Statistiques sur les erreurs moyennes logarithmiques individuelles des espèces en choisissant les coefficients de régularisation avec le critère du minimum pour l'étape C et le critère du *one-standard-error* pour l'étape R, et les seuils avec la méthode séparée.

	Médiane	Moyenne	Maximum
Train	0.049	0.240	1.399
Test	0.078	0.233	1.233

TABLE 3.3 – Statistiques sur les erreurs moyennes logarithmiques individuelles des espèces en choisissant les coefficients de régularisation avec le critère du minimum pour l'étape C et le critère du *one-standard-error* pour l'étape R, et les seuils avec la méthode jointe, avec un seuil de 0.5.

le λ choisi avec le critère du *one-standard-error* retient moins de variables qu'avec le critère du minimum. Comme le LASSO du minimum était déjà proche de prédire la moyenne, alors celui du *one-standard-error* l'est encore plus car il est plus simple.

En affichant les variables qui ont été sélectionnées (figure A.11), on constate que la plupart des modèles n'en n'ont aucune. Ceux-ci prédisent donc la moyenne exactement. L'effet est logiquement plus marqué que précédemment (figure A.10). Un petit nombre de modèles retiennent quelques variables, ce qui fait que le modèle reste légèrement meilleur que B_2^{test}. Enfin, comme on ne possède pas beaucoup d'observations, la taille de l'ensemble test est petite ($n_{test} = 109$). Par chance, il se peut que la moyenne des données de test corresponde bien à la moyenne des données d'entraînement, alors sa variance est inférieure.

Si on se penche sur les résultats de l'étape C avec les métriques habituelles de la classification (table 3.4), on remarque que la sensibilité globale de l'étape C est de seulement 57.5% sur l'ensemble d'entraînement : le LASSO rate beaucoup de valeurs positives. En revanche, la spécificité est très élevée : il réussit très bien à détecter les absences de rejet. De plus, les résultats de test (table 3.5) sont similaires à ceux de l'entraînement et n'indiquent aucun surajustement : les variables sélectionnées par les modèles (figure A.8) sont donc pertinentes. En particulier, parmi les plus utilisées (figure 3.7), on retrouve les variables spatiales (X_{lat} et X_{lon}), les variables temporelles (X_{sin} et X_{cos}), ainsi que les données de débarquement des espèces les plus fréquentes comme X_{89} ou X_{60}, correspondantes à la langoustine commune (*Nephrops norvegicus*) et à la cardine franche (*Lepidorhombus whiffiagonis*). Une analyse plus fine révèle que les débarquements d'une espèce sont souvent utilisés pour prédire la présence de ses rejets.

Ces résultats du modèle à deux étapes sont donc contrastés. Les modèles de régression pénalisées de l'étape R semblent mal s'ajuster aux données, même non-nulles. En revanche, l'étape C semble atteindre son objectif et induit l'essentiel du gain en coût par la bonne prédiction des zéros. Des études futures pourraient se concentrer sur l'exploration d'autres modèles de régression pour l'étape R et sur la réduction des faux négatifs pour l'étape C.

	$\hat{V}_k^{(i)} = 1$	$\hat{V}_k^{(i)} = 0$
$V_k^{(i)} = 1$	2158	1596
$V_k^{(i)} = 0$	494	33702

TABLE 3.4 – Matrice de confusion de l'étape C pour les données d'entraînement. $n_{train} \times K = 253 \times 150 = 37950$, sensibilité = 57.5%, spécificité = 98.6%.

	$\hat{V}_k^{(i)} = 1$	$\hat{V}_k^{(i)} = 0$
$V_k^{(i)} = 1$	945	680
$V_k^{(i)} = 0$	187	14538

TABLE 3.5 – Matrice de confusion de l'étape C pour les données de test. $n_{test} \times K = 109 \times 150 = 16350$, sensibilité = 58.2%, spécificité = 98.7%.

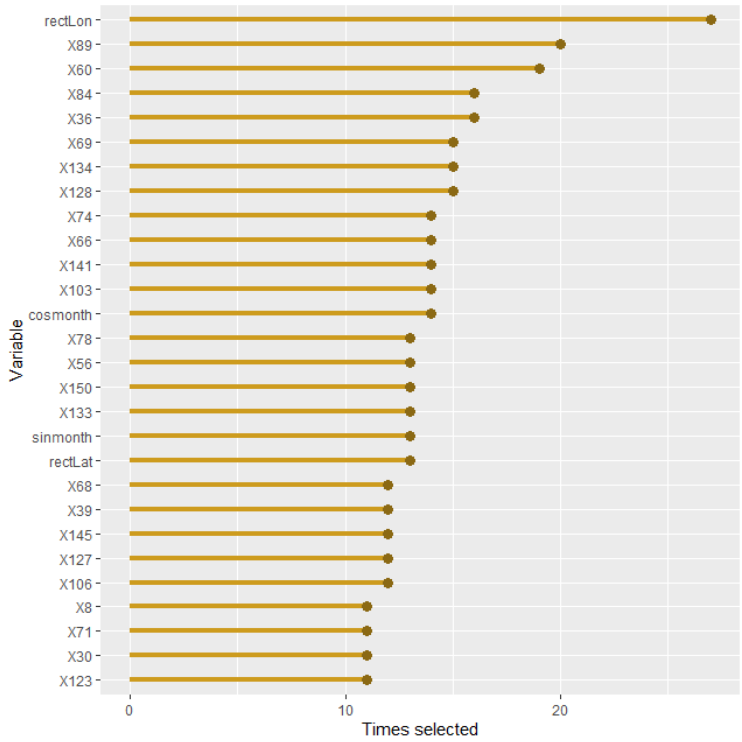

FIGURE 3.7 – Variables sélectionnées dans plus de 10 modèles de l'étape C et le nombre de modèles dans lesquels ils apparaissent.

Méthode appliquée aux modèles de régression antérieurs

A priori, peu d'arguments théoriques permettent de préférer *step* ou la régularisation l_1 pour la sélection de variables (Wasserman & Roeder, 2009 ; Hastie et al., 2017). Il faut parfois se résoudre à essayer plusieurs techniques afin de comparer leurs résultats. On a donc appliqué la procédure *step* mentionnée dans le chapitre précédent au modèle à deux étapes. Comme on le voit sur la figure 3.8, la discontinuité de l'erreur est effacée par le processus C puis R. Les résultats sont très similaires entre les ensembles d'entraînement et de test (table 3.6). De fait seulement 0.9% des coefficients sont non nuls pour le modèle joint (8.5% pour le modèle simple), ce qui rejoint la conclusion de la régularisation l_1. De manière générale, l'erreur est moins bonne que pour la régularisation l_1.

Modèle	Moyenne Entraînement	Moyenne Test	Max Entraînement	Max Test
Prédiction de moyenne	0.39	0.38	3.95	3.96
Régression simple	0.37	0.36	3.95	3.96
Régression jointe	0.39	0.38	3.95	3.96

TABLE 3.6 – Statistiques sur les erreurs moyennes logarithmiques individuelles des espèces obtenues avec différents modèles de régression *step*. Les modèles obtiennent d'après ces indicateurs des performances similaires.

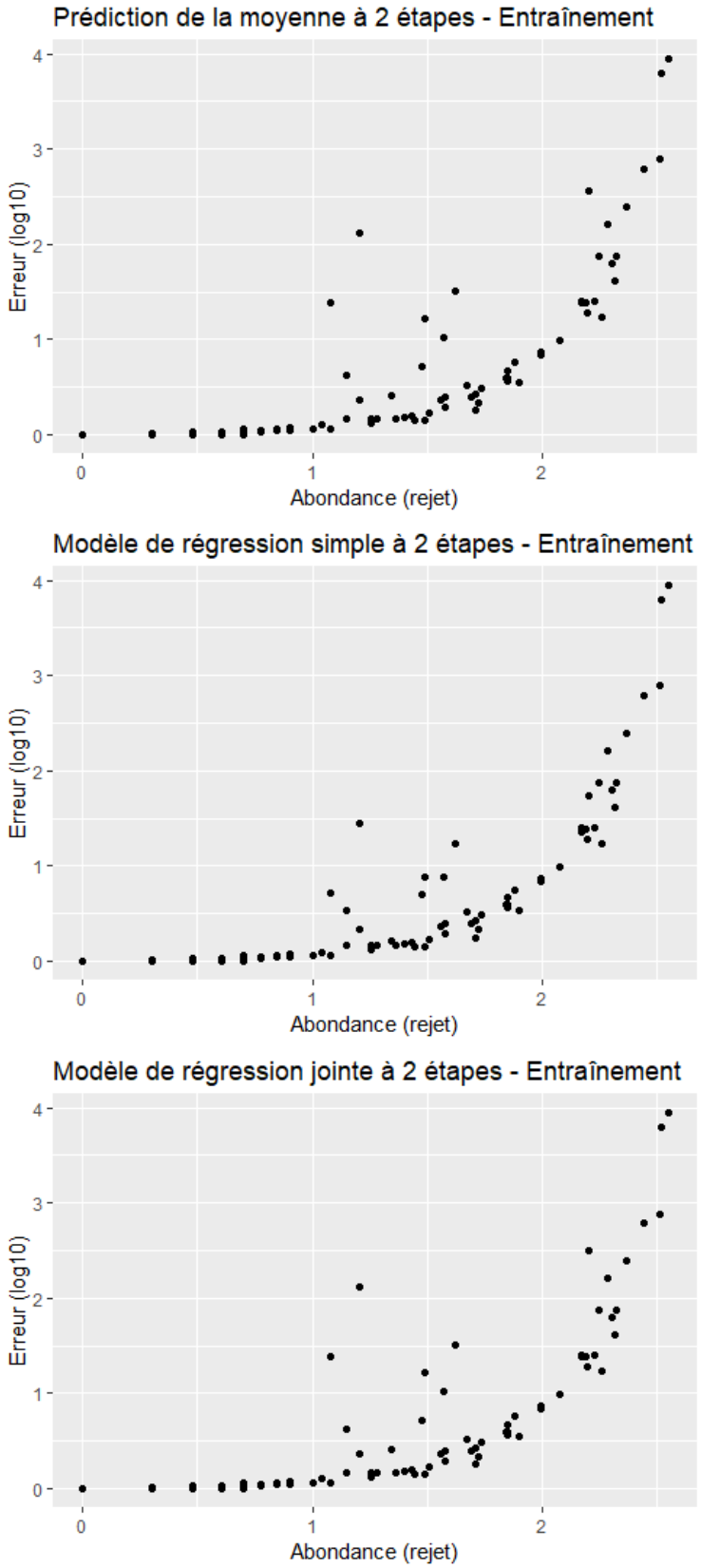

FIGURE 3.8 – Erreur pour l'ensemble d'entraînement en fonction de la fréquence des rejets pour trois modèles (de haut en bas) : prédiction de la moyenne, régression linéaire simple, régression linéaire jointe. Les distributions sont similaires sur l'ensemble de test.

Chapitre 4

Random forest

Les modèles que l'on a considérés jusqu'à présent sont probabilistes et reposent sur des hypothèses fortes. On a pu observer que quand ces dernières n'étaient pas respectées, ils donnaient des prédictions passables. De façon notable, l'excès de zéros et l'hétéroscédasticité ont requis des solutions qui n'étaient pas forcément naturelles et malgré cela, on n'a pas pu obtenir des résultats probants, notamment pour prédire des données positives. On propose d'explorer une méthode non-probabiliste, le modèle des forêts aléatoires (*random forest*), qui repose sur l'agrégation d'arbres de décision.

À nouveau, on ne considère les données de rejet que d'une seule espèce $Y_k =: Y$. On appliquera les modèles qui suivent à chaque Y_k de manière indépendante. Pour estimer les quantités rejetées $\mathbf{Y}$ à partir de l'observation des variables $\mathbf{X} = (\mathbf{X}_j)_{1\leq j\leq p}$, on a besoin d'un régresseur f pour décider $\hat{Y}^{(i)} = f(X^{(i)})$ pour $i = 1...n$.

1 Principe

1.1 Arbre de régression

Pour présenter les arbres de régression, adaptons le formalisme de Hastie et al. (2009) à nos notations. Soit $\mathcal{X}_j$ l'espace de la j-ème variable explicative, typiquement $\mathbb{R}^+$ pour les quantités débarquées, et $\mathcal{X} = \bigoplus_{j=1}^{p} \mathcal{X}_j$ l'espace de l'ensemble des variables explicatives. Si on réalise un partitionnement de cet espace en M parties $(R_m)_{1\leq m\leq M}$, alors on peut attribuer une valeur c_m à chaque. Ainsi, pour tout $x \in \mathcal{X}$, il existe un unique m tel que $x \in R_m$. On peut alors définir un régresseur f de la manière suivante :

$$f(x) = \sum_{m=1}^{M} \mathbb{1}_{x\in R_m} c_m.$$

Le choix des valeurs de $(c_m)_{1\leq m\leq M}$ se fait en optimisant une fonction de coût. Pour le coût des moindres carrés, il est évident que la valeur optimale pour c_m est la moyenne des observations qui sont dans R_m. Le principe d'un arbre de régression est de créer un partitionnement de manière récursive, qui sert ensuite à définir son régresseur.

Un arbre de décision binaire ne peut faire de partition que d'une seule manière. Pour un espace R correspondant à un noeud, il choisit une variable X_j et un seuil s valides, puis il définit R_1 et R_2

ses feuilles telles que :

$$R_1(j,s) = \{X \in R : X_j < s\} \quad \text{et} \quad R_2(j,s) = \{X \in R : X_j \geq s\} \ . \tag{4.1}$$

De manière récursive, on peut obtenir un partitionnement de $\mathcal{X}$ en faisant pousser l'arbre. À l'itération h, $\mathcal{X}$ est partitionnée en h parties $R_1, ..., R_h$, qui correspond aux h noeuds terminaux (feuilles) de l'arbre. L'algorithme (Breiman et al., 1984) choisit :

— une feuille de l'arbre, qui correspond à une partie R_l, $1 \leq l \leq h$,
— un couple (j, s) correspondant à la variable et au seuil pour faire la partition de R_l selon la procédure (4.1).

On fait ainsi pousser deux feuilles à ce noeud de l'arbre, c'est-à-dire deux nouvelles parties. On recommence cette procédure jusqu'à l'arrêt de l'algorithme, que l'on détermine par exemple en définissant un nombre maximal de partitions M ou en imposant un nombre minimal d'observation à avoir dans chaque feuille.

On constate qu'il est nécessaire de définir des critères pour les choix de la feuille à développer R_l et du couple (j, s) pour la partition. Pour le choix de R_l, on explore souvent l'arbre de manière exhaustive tant qu'il y a assez d'observations dans chaque partition ou jusqu'à une profondeur maximale. Pour (j, s), on utilise la plupart du temps un critère glouton : on prend celui qui minimise le coût. Enfin, on peut réduire la complexité de l'arbre en enlevant les noeuds les moins significatifs (*pruning*).

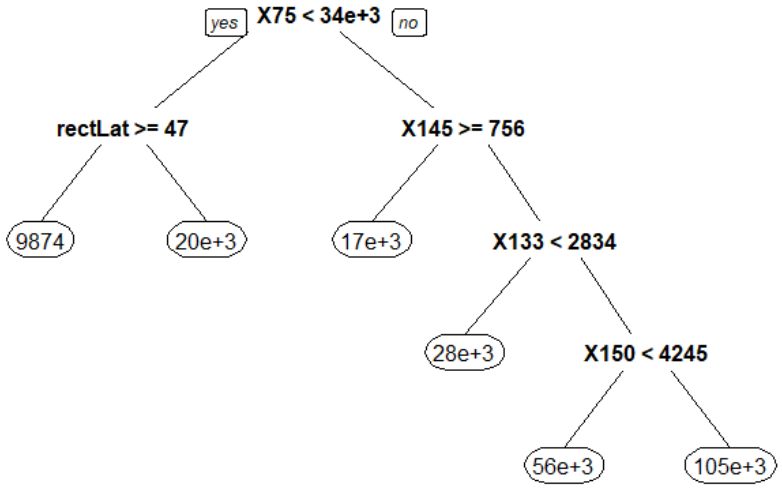

FIGURE 4.1 – Arbre de régression pour Y_{75}, les données de rejet du merlu commun, obtenu à partir d'un arbre quasi-complet (figure A.12) réduit après étude de ses performances par validation croisée 10-fold.

Enfin, la représentation graphique d'un arbre est souvent parlante car elle indique automatiquement les variables explicatives et elle donne une interprétation possible. La figure 4.1 montre un exemple d'arbre de régression construit pour prédire Y_{75}, les quantités débarquées du merlu commun (*Merluccius merluccius*). Sans surprise, on note le rôle important de ses propres quantités de débarquement (X_{75}) et des données géographiques (X_{lat}) dès les premières partitions.

Malgré la flexibilité offerte dans le choix du critère d'arrêt, il n'existe pas de règle de choix optimale a priori. Malgré l'utilisation possible de la validation croisée, l'aspect glouton de l'algorithme de construction de l'arbre l'empêche explorer efficacement l'espace des variables pour trouver les

meilleures partitions, d'autant plus que le choix des partitions futures dépend fortement de celles choisies en amont. On conclut en notant que si on construit un arbre complet sans limitation sur le nombre minimal d'observations, on aura à la fin de l'arbre complet une observation par feuille. On aura donc parfaitement ajusté le modèle aux données d'entraînement : l'erreur est nulle. Par contre, on a évidemment surajusté le modèle et il est probable que l'arbre soit mauvais sur des données nouvelles. Ainsi, même si on peut diminuer leur biais à souhait, les arbres de régression sont une méthode à grande variance.

1.2 Forêts aléatoires

À partir d'un ensemble d'observations de taille n, le *bootstrapping* consiste à échantillonner directement dans cet ensemble pour former des ensembles "bootstraps". Ceux-ci sont constitués en effectuant n tirages aléatoires avec remise dans l'ensemble initial. En quelque sorte, les ensembles bootstraps sont donc des perturbations de l'ensemble d'origine.

Le *bagging* consiste à agréger des estimateurs d'un certain modèle entraînés sur des ensembles bootstraps différents (Breiman, 1996). Il est prouvé que cette technique diminue la variance de l'estimateur agrégé et que pour assez d'estimateurs bootstraps, le biais introduit par cette procédure est négligeable.

Les forêts aléatoires (*random forests*) combinent ces techniques statistiques et les appliquent aux arbres de décision (Breiman, 2001). L'algorithme consiste à produire un grand nombre d'arbres entraînés sur des ensembles bootstraps. De plus, les arbres ne sont plus construits de manière complètement gloutonne comme précédemment, mais en introduisant des perturbations aléatoires dans le choix des partitions. Enfin, le régresseur de la forêt fait la moyenne de la prédiction de tous ses arbres.

En particulier, une forêt comporte deux hyperparamètres B le nombre d'arbres à construire et m le nombre de variables utilisées dans la recherche des partitions. Ces paramètres s'insèrent dans l'algorithme de construction de la forêt de la manière suivante, pour $b = 1...B$:

1. Construction d'un ensemble bootstrap $\mathbf{X}^{(b)}$ de l'ensemble d'entraînement.
2. Construction de l'arbre $T^{(b)} = T(\mathbf{X}^{(b)})$ à partir de $\mathbf{X}^{(b)}$ comme précédemment, sauf qu'à chaque itération, on cherche la meilleure partition (j, s) avec j dans un ensemble de m variables tirées au hasard (au lieu de l'ensemble de toutes les variables).

Pour chaque arbre $T^{(b)}$ de la forêt, on a un régresseur $f_{T^{(b)}}$. Finalement, le régresseur de la forêt s'écrit :

$$f^B(x) = \frac{1}{B} \sum_{b=1}^{B} f_{T^{(b)}}(x).$$

Le rôle des hyperparamètres est donc clair. Augmenter B fait diminuer la variance de f^B et par propriété du *bagging*, cela fait converger f^B sans surajustement. Cependant, le rendement est décroissant et le coût computationnel augmente linéairement avec B. m quantifie la part d'aléatoire dans la construction de l'arbre. Si on choisit $m = p$, le nombre de variables explicatives, alors on construit des arbres comme présenté dans le paragraphe précédent de manière déterministe (mais sur les ensembles bootstraps). Plus on choisit m petit, plus les arbres sont décorrélés, ce qui rend l'agrégation plus efficace. Mais cela augmente aussi le biais de l'ensemble, ce qui peut devenir contre-productif. C'est donc un paramètre à régler empiriquement (Segal, 2004) afin trouver un équilibre entre biais et variance, à la manière d'un coefficient de régularisation.

2 Application

Avec les données ObsMer, on s'attend à ce que les forêts aléatoires donnent des résultats décents. En effet, autant l'excès de zéros et l'hétéroscédasticité ont posé des problèmes difficiles à surmonter dans les modèles probabilistes, ceux-là ne devraient pas trop influencer les forêts aléatoires. Pour les espèces avec un grand nombre de données nulles, on a l'espoir que les arbres trouvent une partition décente pour isoler leurs zéros dans un noeud. Cependant, contrairement au modèle à deux étapes, une forêt aléatoire ne prédira que rarement 0 exactement. En effet, il suffit qu'un arbre seul de la forêt ne trouve pas de partition isolant les zéros pour que la moyenne devienne strictement positive. Toutefois, au lieu de prédire 0, la forêt devrait prédire des valeurs faibles, ce qui reste peut-être acceptable.

2.1 Implémentation

Le processus d'inférence par les forêts aléatoires se présente de la manière suivante :

1. Transformation logarithmique (*ou pas*) des quantités débarquées $\tilde{Y}_k \leftarrow \log_{10}(Y_k + 1)$, pour $k = 1, ..., K$;
2. Création d'un ensemble d'entraînement et d'un ensemble de test ;
3. Ajustement d'une forêt aléatoire sur les données de l'ensemble d'entraînement pour chaque espèce k ;
4. Prédiction sur les données de l'ensemble de test en utilisant la forêt associée à l'espèce pour chaque espèce k ;
5. Calcul des erreurs de prédiction sur toutes les espèces.

La pertinence de transformer les données comme dans les modèles probabilistes est questionnable. En effet, en écologie, il n'y a pas de consensus sur la manière d'appliquer les forêts aléatoires à des données d'abondance. En fonction du problème, une transformation logarithmique (Barber et al., 2016) ou binaire (Hill et al., 2017) est souvent appliquée, mais ce n'est pas systématique (Oppel et al., 2012). Même si l'hétéroscédasticité n'est pas un facteur qui entre en compte cette fois, on choisit tout de même d'effectuer la même transformation logarithmique. En effet, on devrait suivre une procédure similaire au modèle à deux étapes pour entreprendre une comparaison juste.

Il existe un grand nombre de solutions implémentant les forêts aléatoires. La raison en est que les forêts aléatoires sont largement étudiées et qu'elles sont très populaires dans les problèmes de régression et de classification. On a décidé d'utiliser le package `randomForest` (Liaw et al., 2002) qui propose une interface R au code Fortran du concepteur original de l'algorithme (Breiman, 2001). Celui-ci permet de faire pousser simplement des forêts en précisant les hyperparamètres B et m.

Les erreurs de prédiction sont calculées comme précédemment avec la métrique logarithmique. Grâce à la transformation des données, on a une nouvelle fois :

$$J_k^{\log} = \frac{1}{n} \sum_{i=1}^{n} |\hat{\tilde{Y}}_k^{(i)} - \tilde{Y}_k^{(i)}|,$$

$$J^{\log} = \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} |\hat{\tilde{Y}}_k^{(i)} - \tilde{Y}_k^{(i)}|.$$

On garde également l'interprétation de ce coût comme une moyenne géométrique sur les différences d'ordre de grandeur (chapitre 2, section 4.1).

2.2 Résultats

On a entraîné les forêts aléatoires avec des hyperparamètres B et m communs à toutes les espèces. On a cherché les paramètres optimaux selon une recherche par quadrillage (*grid search*). Comme attendu, augmenter B fait baisser de manière monotone les erreurs sur les données d'entraînement. Néanmoins, à partir de $B > 200$, les améliorations ne sont plus claires et les erreurs sur l'ensemble de test stagnent. Comme choisir un plus grand B ne dégrade pas les performances des forêts, on prend un peu de marge en choisissant $B = 500$. En ce qui concerne nos données, un m trop petit mène à des résultats chaotiques et très variables. Par contre, on n'observe pas de différence notable entre tous les $m > 50$. On choisit donc $m = 50$. Sachant que le nombre de variables est $p = 154$, m représente approximativement un tiers d'entre elles.

Les résultats sont conformes aux espérances (table 4.1) et peuvent être comparés à ceux du modèle à deux étapes. Pour l'ensemble d'entraînement, ils sont excellents, avec des erreurs très faibles. Les forêts aléatoires sont donc très peu biaisées. En revanche, on observe que les performances sur les données de test sont moins bonnes que dans le cadre du modèle à deux étapes. Cela témoigne d'un léger surajustement aux données. On rappelle que la procédure de *bagging* des forêts aléatoires ne peut pas induire de surajustement, au contraire des arbres eux-mêmes qui sont construits par cette procédure. On gagnerait donc à réduire la taille des arbres, par exemple en imposant un plus grand nombre minimal d'observations par feuille ou une profondeur maximale.

	Médiane	Moyenne	Maximum
Train	0.049	0.152	0.721
Test	0.093	0.337	1.562

TABLE 4.1 – Statistiques sur les erreurs moyennes logarithmiques individuelles des espèces obtenues avec des forêts aléatoires.

Enfin, il est possible d'obtenir l'importance des variables explicatives dans les forêts. En effet, on peut chercher celles qui induisent les plus grandes augmentations de l'erreur lorsqu'on les détruit en les permutant aléatoirement. La figure 4.2 indique les variables les plus significatives pour chaque forêt (correspondante à une espèce rejetée). Encore une fois, on retrouve les variables spatio-temporelles (X_{lat}, X_{lon}, X_{sin} et X_{cos}) parmi les plus présentes. Le tacaud commun (*Trisopterus luscus*, Y_{145}) admet par exemple une grande dépendance par rapport à la longitude X_{lon}, alors que que le merlan bleu (*Micromesistius poutassou*, Y_{77}) dépend beaucoup de X_{cos}. Comme X_{cos} est le cosinus du mois placé sur le cercle trigonométrique, il met en valeur les différences printemps-été contre automne-hiver (figure 4.3). On retrouve aussi la tendance des données de débarquement d'une espèce à jouer un rôle important dans la prédiction de ses propres rejets. Pour citer des espèces différentes que dans la section précédente, mentionnons la baudroie rousse (*Lophius budegassa*, Y_{39}) et le grondin gris (*Eutrigla gurnardus*, Y_{69}).

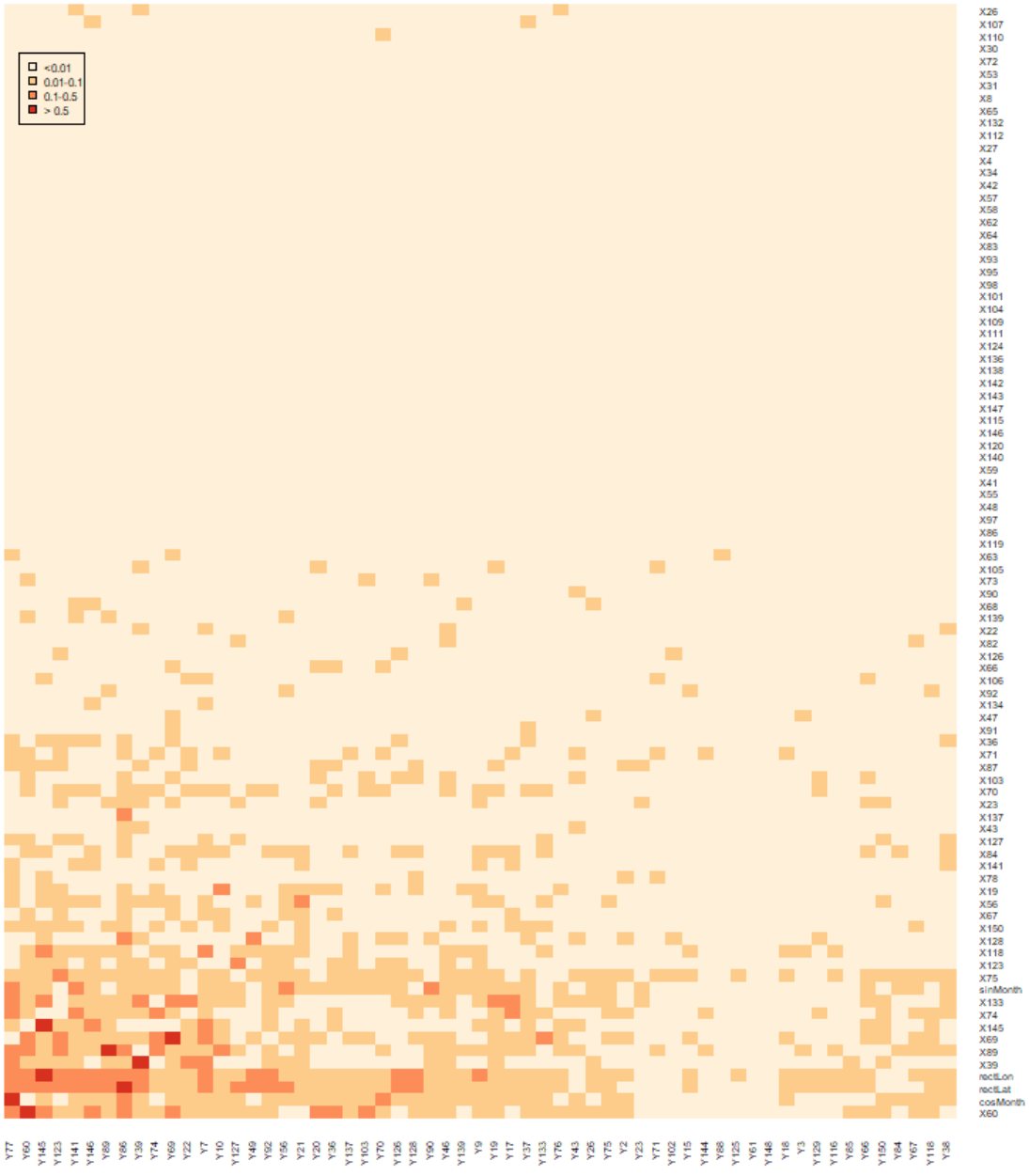

FIGURE 4.2 – Importance des variables dans les forêts aléatoires de chaque espèce débarquée. La métrique utilisée est l'augmentation de l'erreur lorsque la variable étudiée subit une permutation aléatoire, en pourcentage.

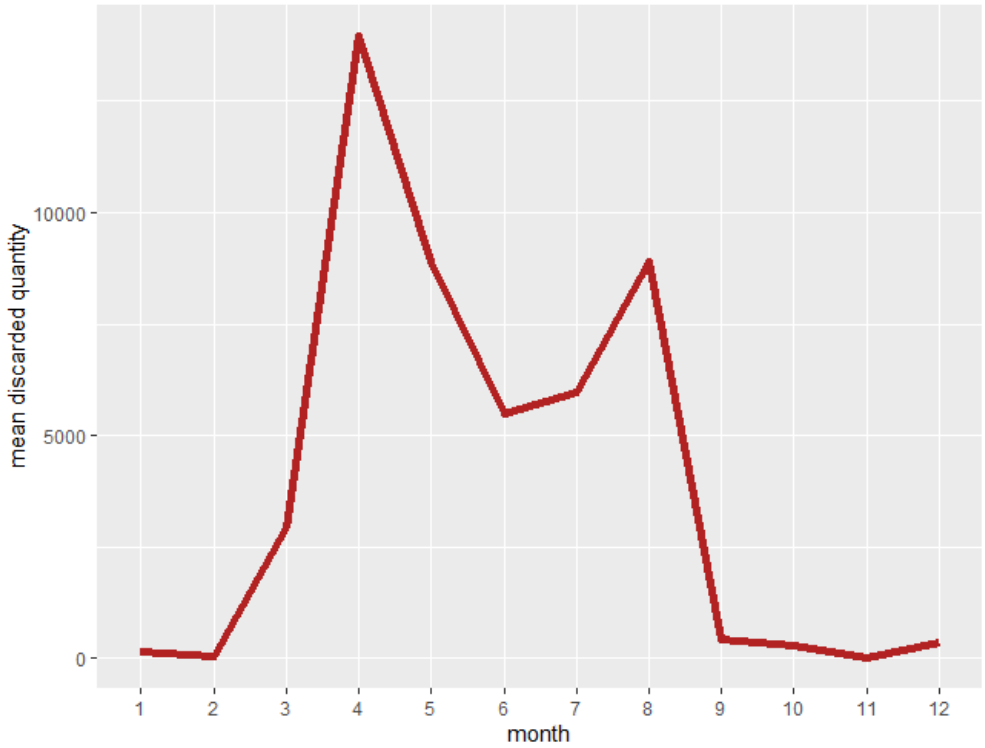

FIGURE 4.3 – Quantités moyennes rejetées pour le merlan bleu (*Micromesistius poutassou*, Y_{77}) en fonction du mois. On constate de grandes variations saisonnières. En particulier, il est très souvent rejeté d'avril à août ($M = 4, .., 8$), c'est-à-dire quand $X_{cos} = \cos(M\pi/6) < 0$.

Chapitre 5

Discussion

1 Difficultés identifiées

Le problème que nous avons traité n'est pas trivial et nous avons pu constater en début de ce rapport que les modèles les plus simples n'étaient pas suffisants. Il faut noter qu'une difficulté majeure est que l'on doive avoir une approche multivariée et non pas traiter les espèces individuellement. Cela est d'autant plus compliqué que les modèles standards impliquant des répartitions gaussiennes ne sont pas exploitables directement.

Par ailleurs, le fait que le jeu de données soit de taille modeste rend les inférences plus sensibles. Notamment, nous avons très peu de données pour les espèces que nous avons considérées comme rares, c'est-à-dire pour celles où moins de 5 rejets ont été observés. Dans ce cadre nous avons choisi de prédire des absences systématiques pour environ 90 espèces sur 150 : une extension des données disponibles permettrait d'incorporer plus d'espèces dans les modèles et donc de proposer des résultats plus intéressants.

2 Pistes d'amélioration

2.1 Représentativité des données

Normalement, les ensembles d'entraînement et de test doivent être représentatifs de la distribution globale des données. Comme nous l'avons évoqué en décrivant la méthode de validation croisée, la construction de l'ensemble d'entraînement est faite de manière naïve. Si cela est a priori sans grande conséquence pour les espèces abondantes, il est probable que l'impact est très élevé sur les prédictions des espèces rares puisqu'on tentera ou non de prédire quelques rejets positifs en fonction de l'aléa de la partition. Il serait aussi intéressant de veiller à bien répartir les opérations de pêche selon les différentes informations spatiales et temporelles.

2.2 Choix de la métrique

Nous avons retenu une métrique logarithmique afin d'empêcher les grandes erreurs d'écraser l'ensemble des prédictions, puisqu'il était admis que nos premiers résultats seraient passables. Cependant, au fur et à mesure du rapport, nous avons obtenu des estimations plus satisfaisantes et il pourrait être nécessaire de choisir une métrique moins laxiste pour mieux distinguer les réelles différences entre modèles.

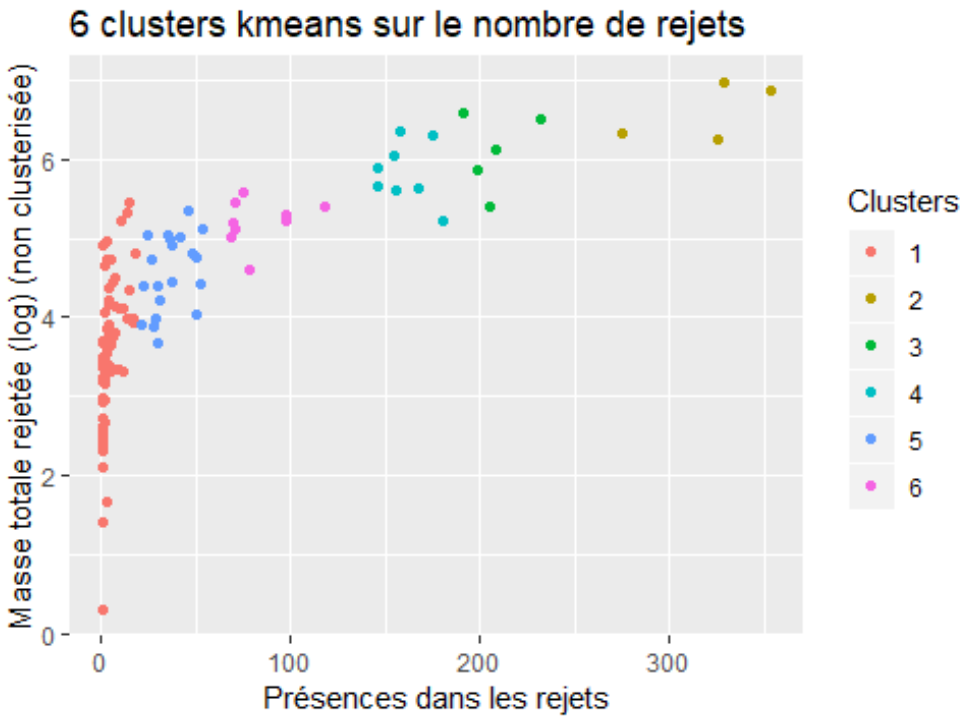

FIGURE 5.1 – Les 120 espèces rejetées au moins une fois sont agrégées (méthode *K-means*) selon leur nombre de rejets dans notre jeu de données. Nous nous intéresserons uniquement au cluster des espèces rares, le numéro 1. La masse rejetée en ordonnée n'est représentée que pour mieux illustrer les données.

2.3 Clusters d'espèces

Il peut être intéressant d'agréger certaines espèces afin de simplifier le modèle en regroupant l'information, en particulier pour les espèces rares. Ces clusters peuvent être établis sur les données à prédire $(Y_k)_{1\leq k\leq K}$ ou les données explicatives $(X_k)_{1\leq k\leq K}$, voire les deux. Le regroupement des variables explicatives crée beaucoup de possibilités, notamment de réduire leur nombre ou d'utiliser des méthodes telles que le *group-lasso* (Yuan & Lin, 2006). C'est une piste de réflexion pour l'avenir, car nous avons plutôt exploré la première idée : en regroupant les espèces rares comparables entre elles, on choisit un seul set de paramètres de régression pour elles toutes, qui peut ainsi être sélectionné à partir de plus de données d'entraînement et ainsi conférer une plus grande robustesse au modèle.

K-means

Une première méthode proposée est celle des *K-means*. Cette méthode construit un nombre donné de clusters selon une grandeur globale de manière à minimiser la distance de chaque point au centre de son cluster. Nous avons tout d'abord effectué un premier clustering des espèces selon leur nombre de rejets. Nous avons retenu 6 clusters en faisant un compromis entre la simplicité du modèle et la valeur des résidus. Le résultat est présenté sur la figure 5.1.

Le cluster qui nous a alors intéressé est celui des espèces rares (cluster 1 sur la figure 5.1), comportant 74 espèces rejetées au moins une fois et avec au plus 18 rejets. Ces 74 espèces ont ensuite été clusterisées en fonction du logarithme de la quantité moyenne rejetée, pour regrouper des espèces ayant des valeurs comparables. L'analyse de l'erreur nous a conduit à choisir 5 clusters d'espèces rares comme on le voit sur la figure 5.2. Au vu des premiers résultats, nous avons écarté

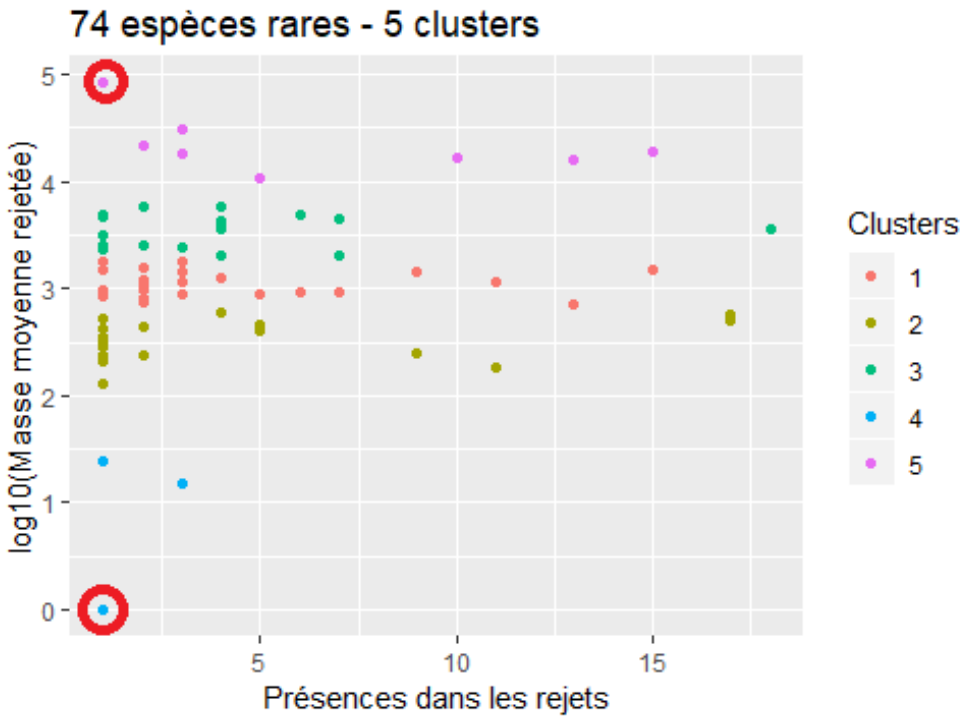

FIGURE 5.2 – Les 74 espèces rares sont cette fois agrégées (méthode *K-means*) selon la quantité moyenne rejetée à chacun de leurs rejets. 5 clusters ont été retenus. La raie brunette et le gobie à quatre taches ont été écartés (entourés en rouge en haut et en bas).

deux espèces singulières : la raie brunette (*Raja undulata*), avec un seul rejet de près de 84 tonnes (très certainement une donnée aberrante), ce qui est en fait très supérieur aux rejets des autres espèces rares de son cluster, ainsi que le gobie à quatre taches (*Deltentosteus quadrimaculatus*) rejeté une fois avec un kilogramme seulement.

On effectue alors les analyses avec 53 "nouvelles espèces" : 46 abondantes, 5 clusters d'espèces rares et les deux espèces rares singulières. Les quantités des clusters sont ici la somme des quantités de toutes les espèces dudit cluster. Les résultats ont semblé en l'état peu prometteurs : pour le modèle de régression simple, la moyenne de l'erreur était par exemple de 0.81 contre 0.52 dans le cas sans clusters. Le problème est que si les espèces agrégées ont des données similaires, il n'y a pas de raison qu'elles aient le même comportement : il faudrait pour cela connaître la corrélation entre leurs présences, ce que ne permet pas le *K-means* qui traite des données globales.

Clusters hiérarchiques

Une seconde méthode de clustering a donc été envisagée : celle des clusters hiérarchiques. Elle repose sur la proximité entre les différentes espèces à partir de toutes les données brutes (soit présence/absence, soit les quantités). En particulier, la corrélation entre les présences dans les différentes opérations de pêche est prise en compte. 72 espèces ont été retenues pour cette analyse : les deux espèces singulières évoquées plus haut ont été retirées.

Nous avons calculé les matrices de proximité entre la masse moyenne des rejets des espèces avec la fonction **dist** et utilisé la fonction **hclust** pour calculer les relations entre elles. Le choix des méthodes retenues est important car, par défaut, celui-ci a tendance à tracer un dendogramme quasi-horizontal, c'est-à-dire qu'il ne parvient pas à séparer les espèces dans différents groupes non-triviaux (presque toutes les espèces sont séparées ou regroupées dans le même cluster). Un

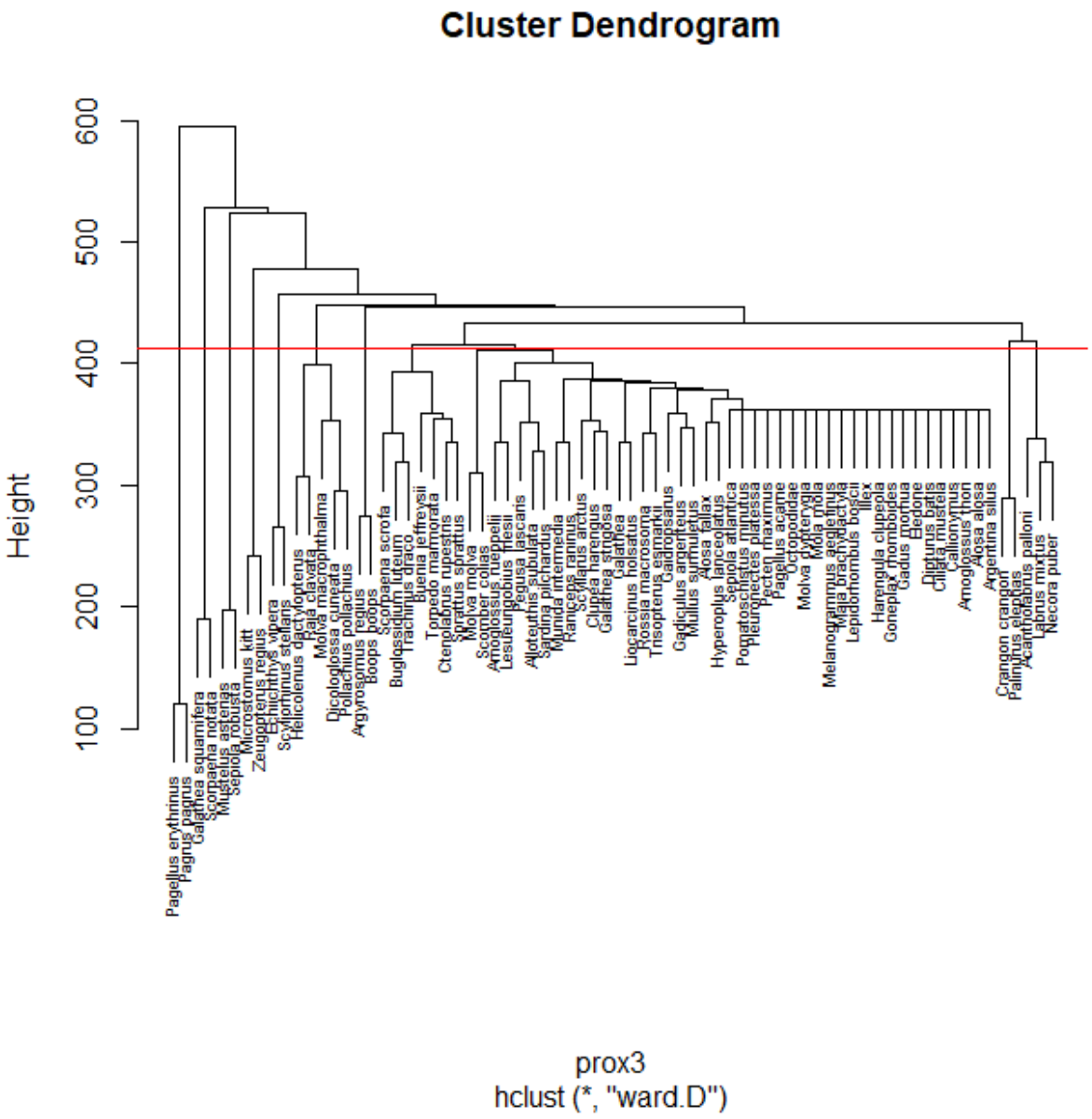

FIGURE 5.3 – Dendogramme pour 72 espèces rares. Les méthodes utilisées pour les matrices de distance, la formation des clusters et la hauteur de coupe ont été choisies "visuellement". 11 clusters sont retenus.

choix satisfaisant est celui de la distance de Canberra (Lance & Williams, 1966) combinée avec une agrégation par la méthode de Ward (`method = "ward.D"`) (Ward Jr, 1963). Il faudrait discuter de la pertinence de ces métriques mais le résultat est visuellement intéressant. La hauteur de coupe choisie pour le dendogramme implique 11 clusters dont 7 sont seulement des paires (et un triplet), comme on l'observe sur la figure 5.3.

L'intérêt est de créer artificiellement plus d'occurrences pour les espèces rares, en considérant que dans un même cluster, chaque espèce est une réalisation indépendante d'un même processus. Si ledit cluster est de cardinal k_{cl}, on observerait donc $n \times k_{cl} = 362 \times k_{cl}$ opérations de pêche pour calibrer les modèles. L'indépendance de ces répétitions ne semble pas garantie. On applique cette procédure pour le modèle de régression linéaire simple. Dans ces conditions, 104 espèces peuvent être traitées avec des prédictions > 0 contre 62 sans les clusters. On obtient des résultats encourageants,

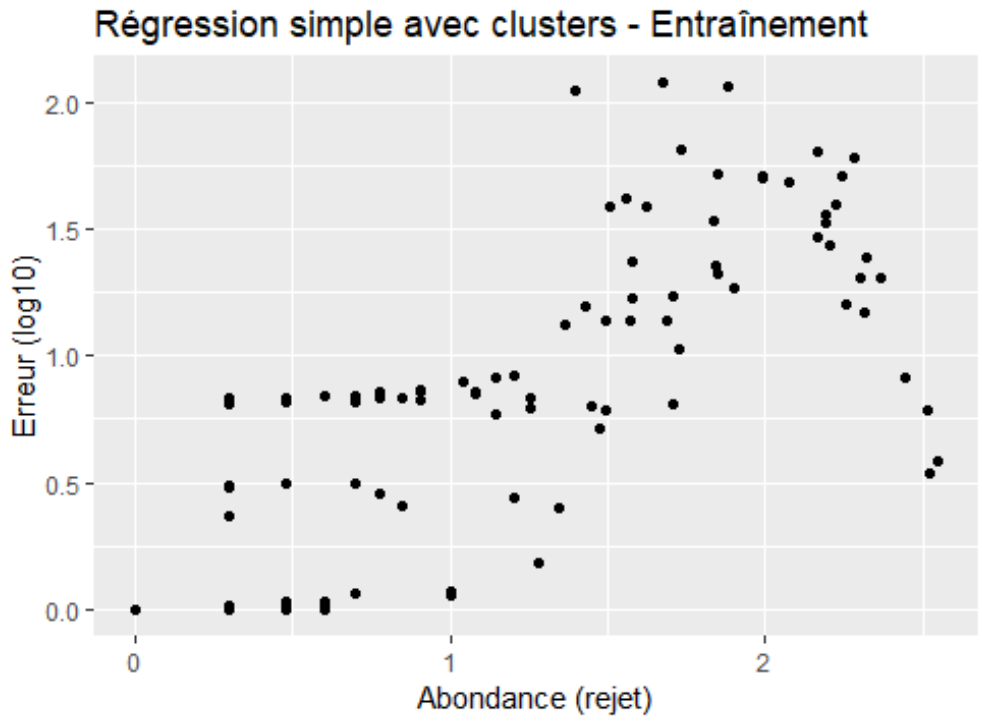

FIGURE 5.4 – Erreur J_k^{log} pour le modèle de régression linéaire simple avec sélection de variables AIC sur les données d'entraînement, avec utilisation de 11 clusters pour les espèces rares. 4 clusters ont permis de prendre en compte 42 espèces rares supplémentaires.

comme on le voit sur la figure 5.4 et dans la table 5.1. En effet, même si l'erreur est moins bonne que sans les clusters (table 2.1), on prend ici un risque plus élevé en prédisant des valeurs non constamment nulles pour plus d'espèces.

Ensemble	Médiane	Moyenne	Maximum
Entraînement	0.81	0.68	2.1
Test	0.86	0.72	2.5

TABLE 5.1 – Répartition de l'erreur J_k^{log} pour le modèle de régression linéaire simple avec sélection de variables AIC. Les 150 espèces sont prises en compte. En moyenne, il y a un rapport $10^{0.7} \simeq 5$ entre les données et l'estimation.

Conclusion

Comme nous l'avons présenté dans ce rapport, nous avons été capables d'obtenir des premières estimations des rejets de pêche à partir des données fournies par l'action ObsMer pour un métier de pêche en particulier, le chalutier langoustinier. Nos travaux constituent essentiellement une approche exploratoire du problème, avec des modèles certes imparfaits mais qui ont le mérite d'avoir révélé les grands enjeux du cas que nous étudions, comme l'impact de la répartition singulière des données avec un grand excès de zéros, de la forte corrélation entre les variables en jeu ou de la grande hétérogénéité des valeurs en termes de quantité et de fréquence.

Si les modèles les plus simples n'ont pas été convaincants, le modèle à deux étapes est lui conceptuellement satisfaisant et nous avons montré que ses résultats étaient prometteurs, en particulier pour l'étape de classification. L'étape de régression n'est pas encore aussi aboutie mais nous avons pu montrer en quoi elle est perfectible, notamment par le fait que peu de variables explicatives sont exploitables par le modèle. De plus, la régression linéaire employée ici a l'avantage d'être un modèle très simple et donc potentiellement plus robuste que des modèles plus ajustés, ce qui est intéressant du fait de la quantité limitée de données à laquelle nous avons eu accès. Par ailleurs, l'implémentation du modèle a été conçue de sorte à ce que des travaux ultérieurs puissent aisément remplacer les modèles de classification ou de régression par d'autres plus pertinents, ce qui ouvre notre projet à des développements potentiels.

Les prédictions par forêts aléatoires ont aussi montré que d'autres approches d'estimation pouvaient être employées avec succès. Nous avons enfin identifié un certain nombre de pistes pour aller plus loin. En particulier, le prochain défi sera de traiter des espèces rares. En l'état actuel et pour un grand nombre des espèces étudiées (de l'ordre de 90 sur 150) nous ne sommes pas en mesure de faire des prédictions non-nulles raisonnables. L'accès à plus de données ou l'agrégation d'espèces pourraient être une solution pour apporter une réponse à cette limite statistique. Il sera intéressant d'évaluer les conséquences des imprécisions de prédiction des rejets pour le groupe d'espèces rares d'un point de vue halieutique. Par exemple pour les espèces rares dont la biomasse est évaluée par des modèles halieutiques, comment cela influence-t-il les évaluations ? Ou encore, pour les espèces rares non-évaluées mais inscrites sur la liste des espèces menacées par l'IUCN, est-ce nécessaire d'estimer la quantité de rejet ou une indicatrice par mois ou par année ne serait-elle pas suffisante ?

Pour rédiger ce rapport, nous avons essayé de trouver une trame narrative dans le travail que nous avons effectué pendant plusieurs mois. Pour cela, nous avons soigneusement sélectionné et réorganisé les pistes les plus prometteuses parmi les nombreuses que nous avons empruntées. Néanmoins, il serait dommage que la trace de certains axes de réflexion soit perdue à tout jamais. De la même manière que nous avons profité des travaux antérieurs, nous espérons que nos éventuels successeurs n'aient pas à tomber dans les mêmes écueils que nous, et s'ils le souhaitent, qu'ils puissent étendre l'exploration au-delà de ce que nous imaginons.

Nous avons donc prêté une attention particulière à ce que notre travail puisse être repris, non seulement grâce à ce rapport, mais également et surtout avec les scripts que nous avons utilisés tout du long. Dans la conception de notre code, nous avons opté pour une structure modulaire. En

guise d'exemple, pour le formatage des données, nous avons un script unique qui les rend utilisables par la plupart de nos modèles, qu'ils soient probabilistes ou pas, qu'ils traitent d'un problème de régression ou de classification. De façon analogue, nos modèles ont également été construits dans cet esprit là. L'architecture du modèle à deux étapes illustre parfaitement notre philosophie. En effet, elle autorise l'utilisateur à définir lui-même un module personnalisé pour l'étape C ou pour l'étape R et de l'intégrer dans le modèle juste par un appel de script séparé. Il évite alors de devoir modifier la structure du modèle à deux étapes.

Nous fournissons également d'autres modèles que nous avons explorés, lesquels n'ont finalement pas été retenus dans le montage définitif du rapport. Citons parmi eux, le *relaxed lasso*, les modèles graphiques avec copules gaussiennes, plusieurs approches de clustering et des analyses exploratoires diverses et variées. Nous implorons cependant nos successeurs à être plus indulgents sur quelques parties de nos scripts qui pourraient sembler relever du bricolage. Si nous nous sommes bien améliorés en R depuis le début de l'année, ce n'est que parce qu'on était au début de nos courbes d'apprentissage respectives, et nous avons encore une grande marge de progression à ce jour.

Enfin, nous tenons à remercier chaleureusement nos encadrantes et encadrant qui nous ont conseillés et suivis au cours des six derniers mois, et espérons que nos travaux leur auront été aussi utiles qu'à nous en tant qu'application des concepts mathématiques étudiés au cours de l'année. Nous remercions également les personnes qui ont effectué le recueil et le *pre-processing* des données.

Références

Barber, Q. E., Bater, C. W., Braid, A. C., Coops, N. C., Tompalski, P., & Nielsen, S. E. (2016). Airborne laser scanning for modelling understory shrub abundance and productivity. *Forest Ecology and Management*, *377*, 46–54.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth.

Catchpole, T., Ribeiro-Santos, A., Mangi, S., Hedley, C., & Gray, T. (2017, 08). The challenges of the landing obligation in eu fisheries. *Marine Policy*, *82*, 76-86. doi: 10.1016/j.marpol.2017.05.001

Commission Européenne. (2013). *Discarding and the landing obligation.* (https://ec.europa.eu/fisheries/cfp/fishing_rules/discards_fr, consulté le 06/03/2020)

Cornou, A.-S., Goascoz, N., Scavinner, M., Chassanite, A., Dubroca, L., & Rochet, M.-J. (2017). *Captures et rejets des métiers de pêche français. résultats des observations à bord des navires de pêche professionnelle en 2016* (Rapport technique). Ifremer. Consulté sur https://archimer.ifremer.fr/doc/00418/52945/

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica : Journal of the Econometric Society*, 829–844.

Cunningham, R. B., & Lindenmayer, D. B. (2005). Modeling count data of rare species : some statistical issues. *Ecology*, *86*(5), 1135–1142.

Devroye, L., Györfi, L., & Lugosi, G. (2013). *A probabilistic theory of pattern recognition* (Vol. 31). Springer Science & Business Media.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv :1707.08692*.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*(5), 531–547.

Hill, L., Hector, A., Hemery, G., Smart, S., Tanadini, M., & Brown, N. (2017). Abundance distributions for tree species in great britain : A two-stage approach to modeling abundance using species distribution modeling and random forest. *Ecology and evolution*, *7*(4), 1043–1056.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Lance, G. N., & Williams, W. T. (1966, 05). Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses"). *The Computer Journal*, *9*(1), 60-64. Consulté sur https://doi.org/10.1093/comjnl/9.1.60 doi: 10.1093/comjnl/9.1.60

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., ... Possingham, H. P. (2005). Zero tolerance ecology : improving ecological inference by modelling the source of zero observations. *Ecology letters*, *8*(11), 1235–1246.

McDavid, A., Gottardo, R., Simon, N., & Drton, M. (2019). Graphical models for zero-inflated single cell gene expression. *The annals of applied statistics*, *13*(2), 848.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, *34*(3), 1436–1462. Consulté sur https://doi.org/10.1214/009053606000000281 doi: 10.1214/009053606000000281

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A. F., Miller, P. I., & Louzao, M. (2012). Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation*, *156*, 94–104.

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression* (Rapport technique). University of California.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, *37*(5A), 2178.

Welsh, A. H., Cunningham, R. B., Donnelly, C., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species : statistical models for counts with extra zeros. *Ecological Modelling*, *88*(1-3), 297–308.

Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, *89*(10), 2953–2959.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *68*(1), 49–67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society : series B (statistical methodology)*, *67*(2), 301–320.

Annexe A

Figures supplémentaires

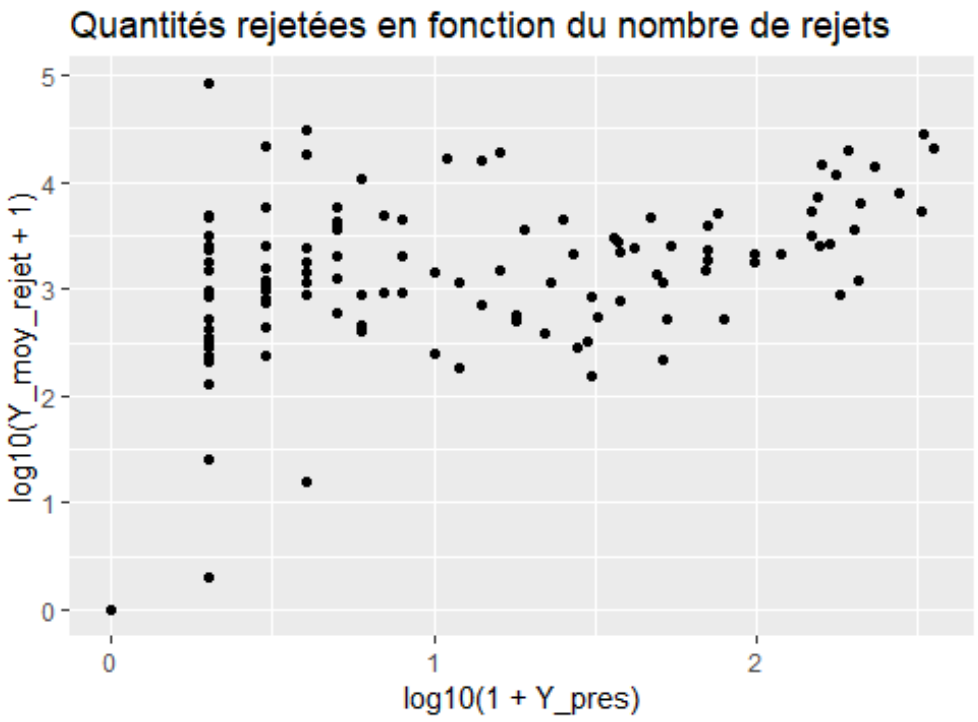

FIGURE A.1 – Relation entre le nombre de rejets et la quantité moyenne des rejets lorsqu'ils ont lieu. Les deux grandeurs ont subi une transformation logarithmique. Chaque point correspond à une espèce.

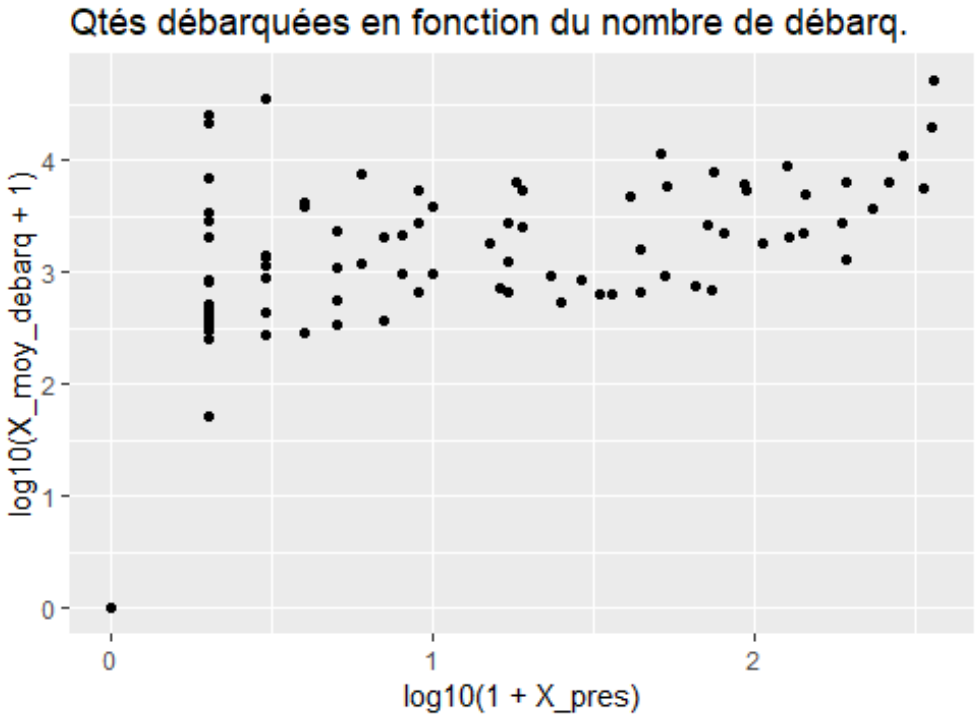

FIGURE A.2 – Relation entre le nombre de débarquements et la quantité moyenne des débarquements lorsqu'ils ont lieu. Les deux grandeurs ont subi une transformation logarithmique. Chaque point correspond à une espèce.

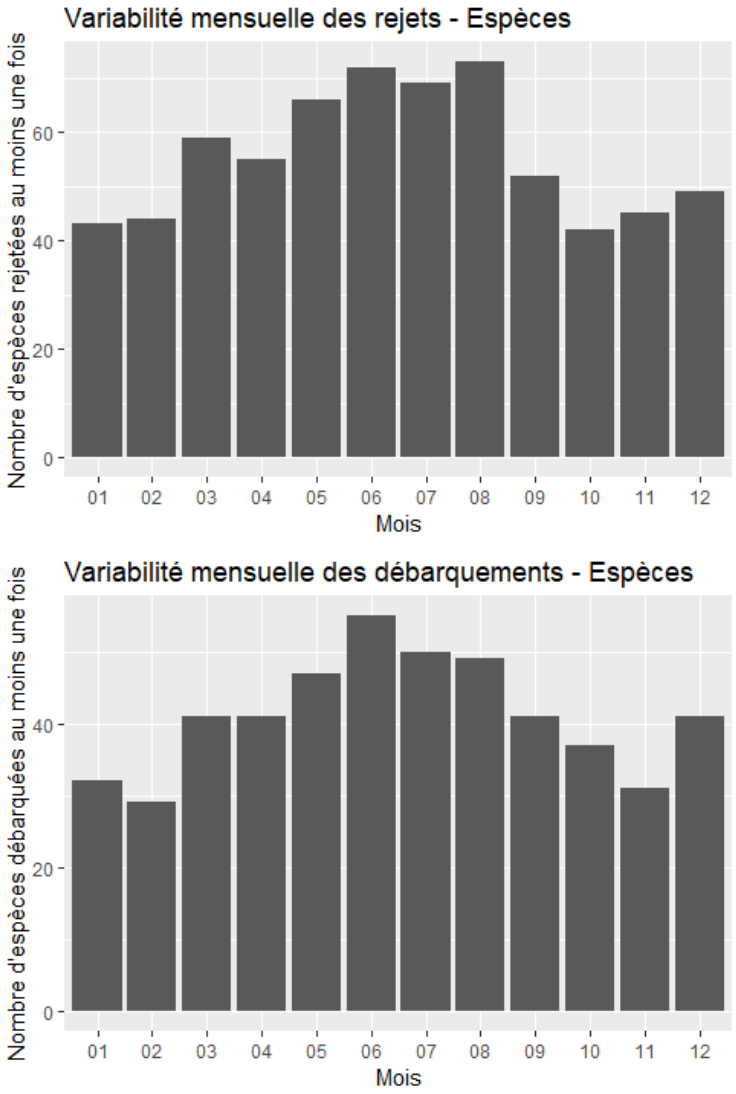

FIGURE A.3 – Nombre d'espèces rejetées (en haut) ou débarquées (en bas) au moins une fois pour chaque mois. On observe une variabilité alors même que ce sont des données agrégées qui ne reflètent pas forcément la spécificité de chaque espèce.

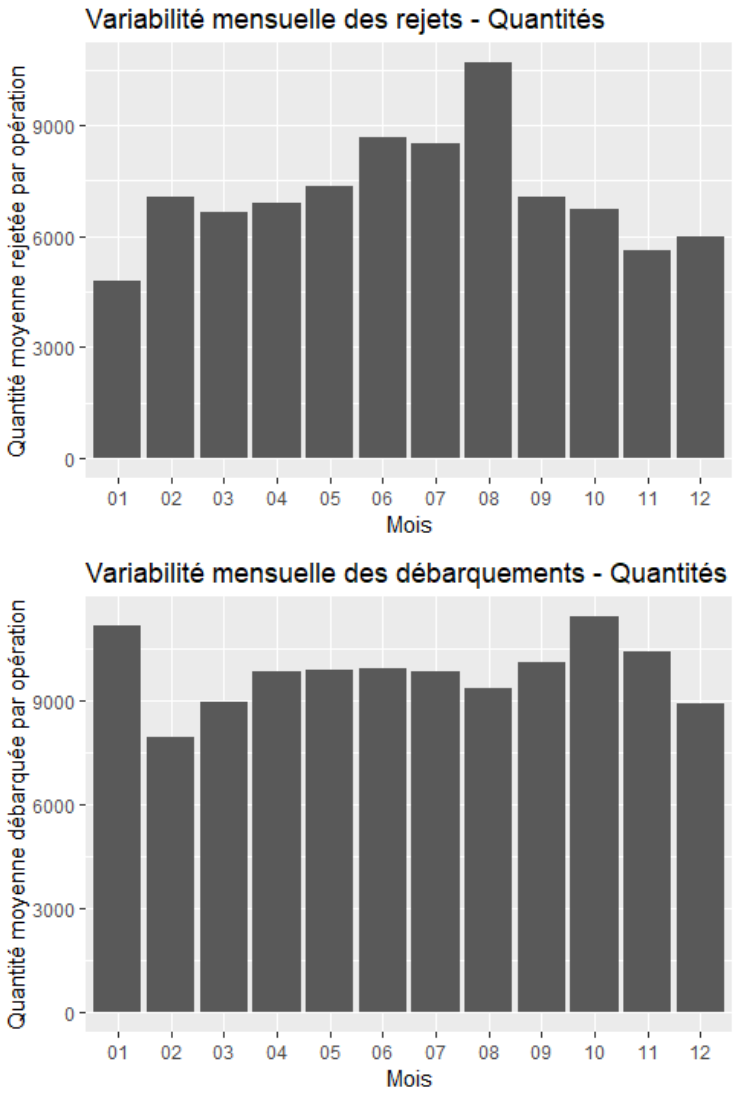

FIGURE A.4 – Quantité moyenne globale rejetée (en haut) ou débarquée (en bas) en une opération en fonction du mois.

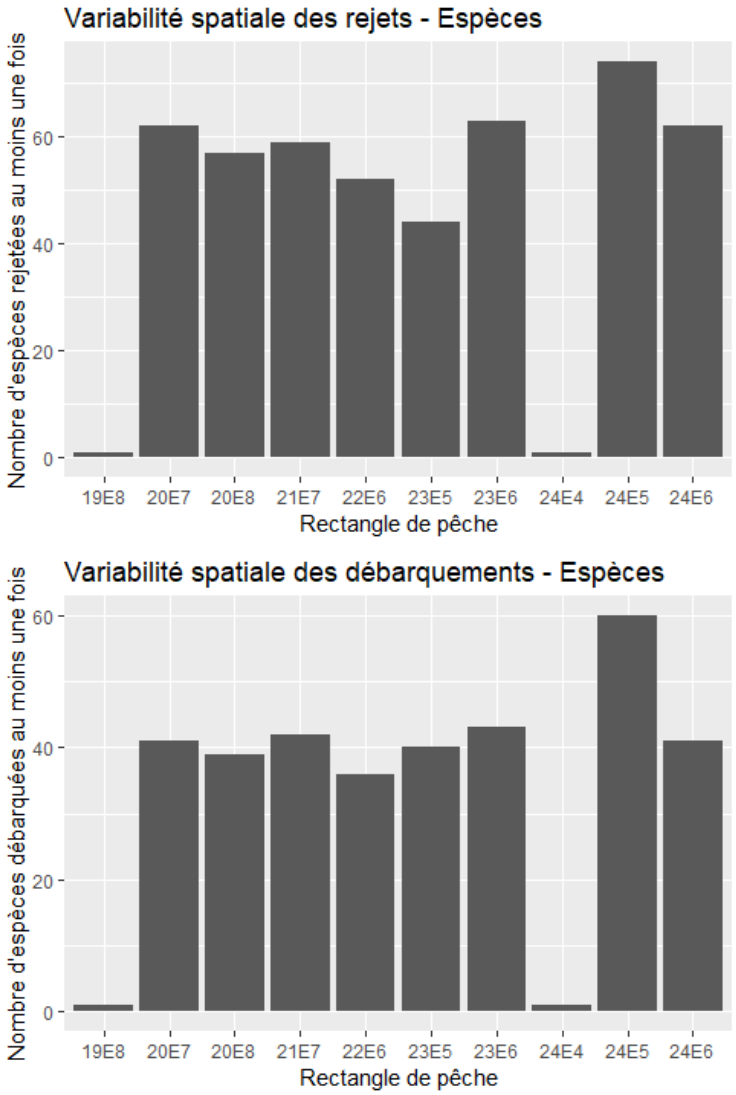

FIGURE A.5 – Nombre d'espèces rejetées (en haut) ou débarquées (en bas) au moins une fois pour chaque rectangle de pêche. On observe une variabilité alors même que ce sont des données agrégées qui ne reflètent pas forcément la spécificité de chaque espèce.

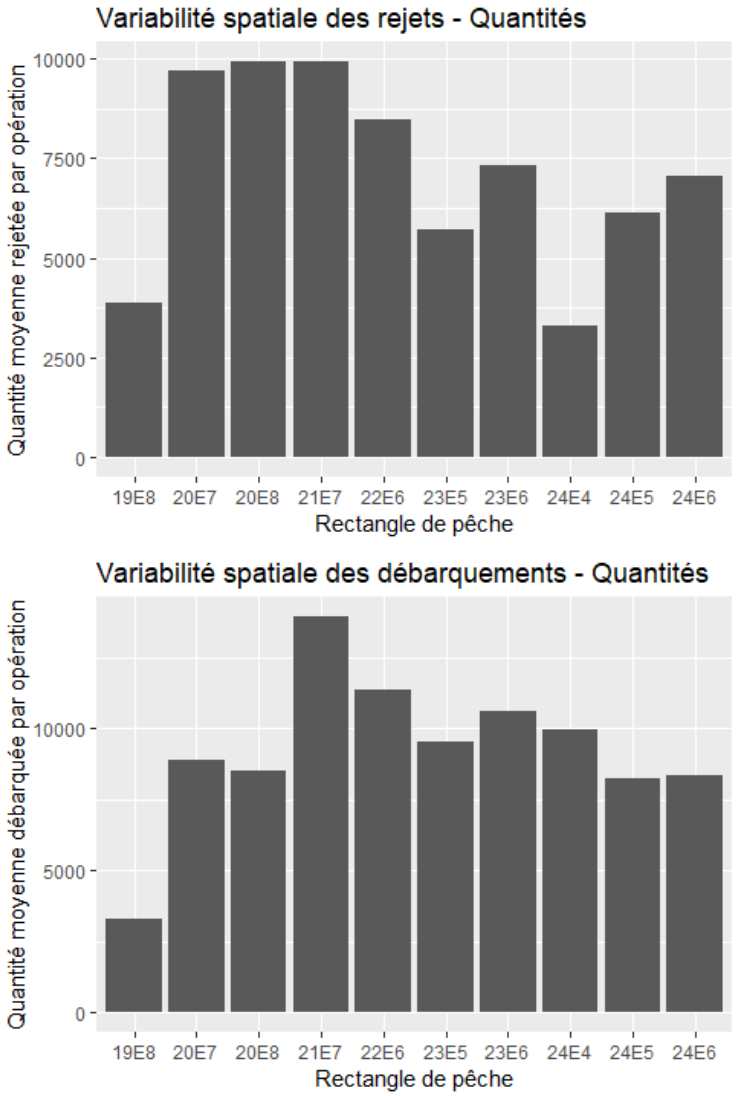

FIGURE A.6 – Quantité moyenne globale rejetée (en haut) ou débarquée (en bas) en une opération en fonction du rectangle de pêche.

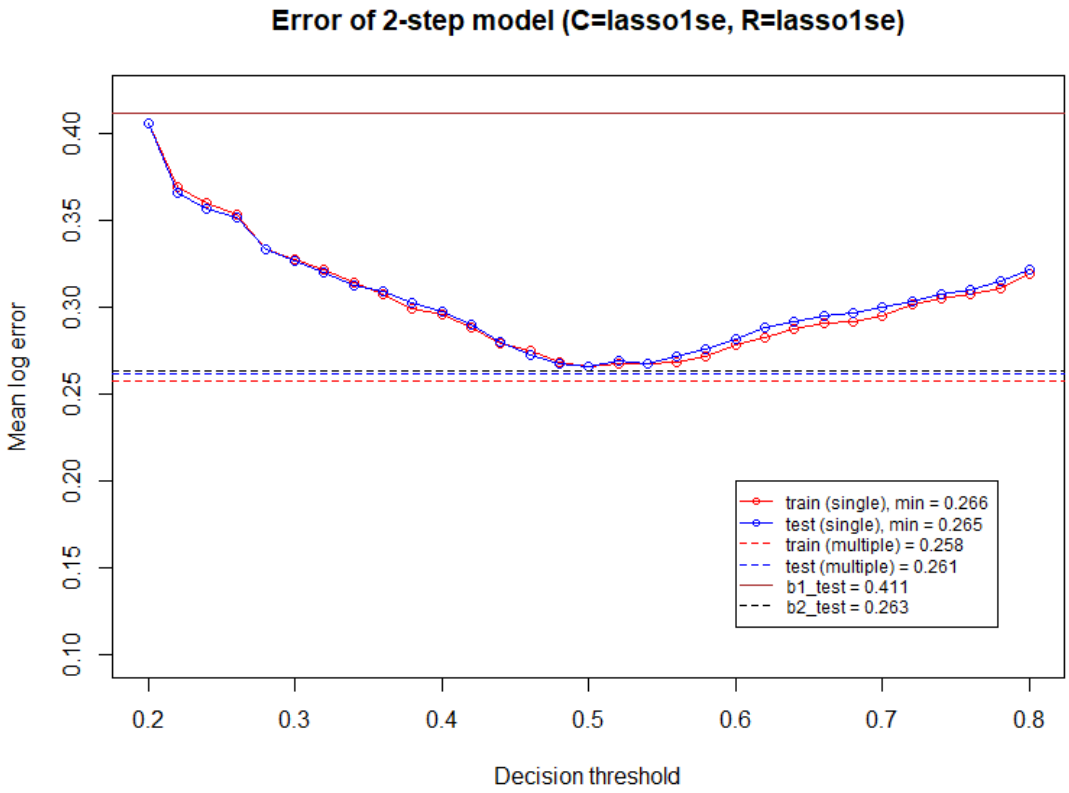

FIGURE A.7 – Erreurs moyennes logarithmiques en choisissant les coefficients de régularisation avec le critère du "one-standard-error" pour les étapes C et R.

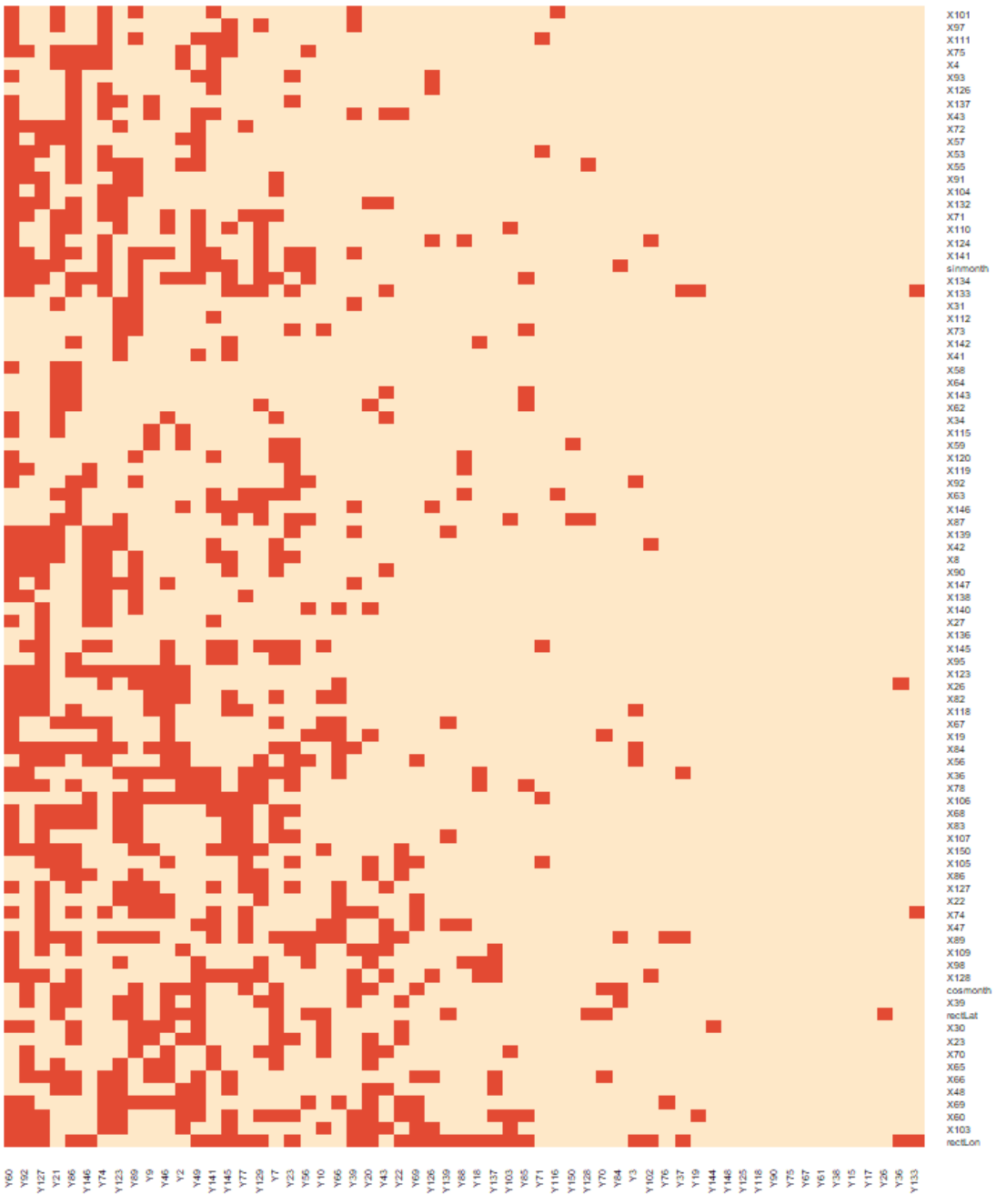

FIGURE A.8 – Variables sélectionnées (en rouge) pour l'étape C en choisissant les coefficients de régularisation avec le critère du minimum.

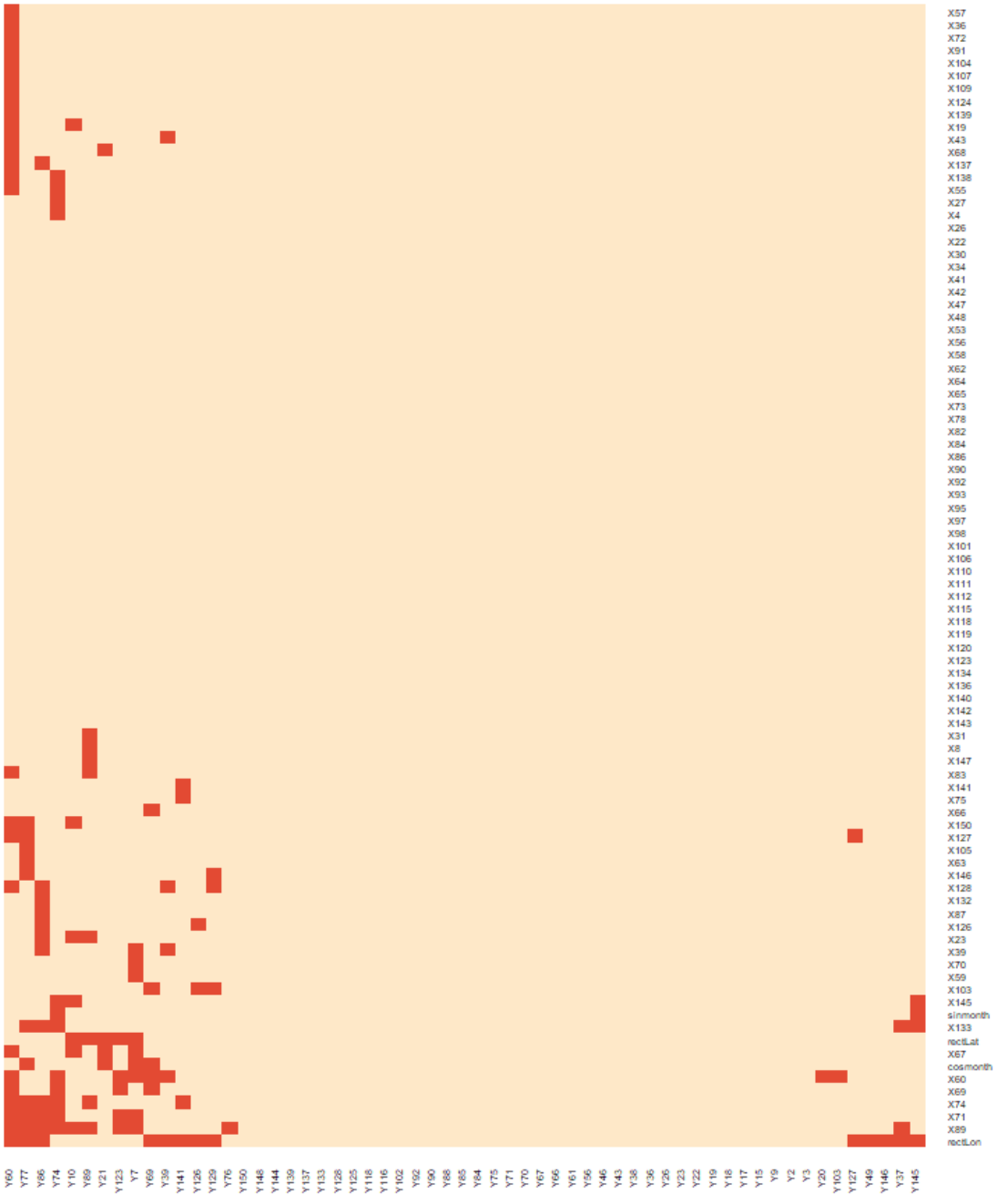

FIGURE A.9 – Variables sélectionnées (en rouge) pour l'étape C en choisissant les coefficients de régularisation avec le critère du "one-standard-error".

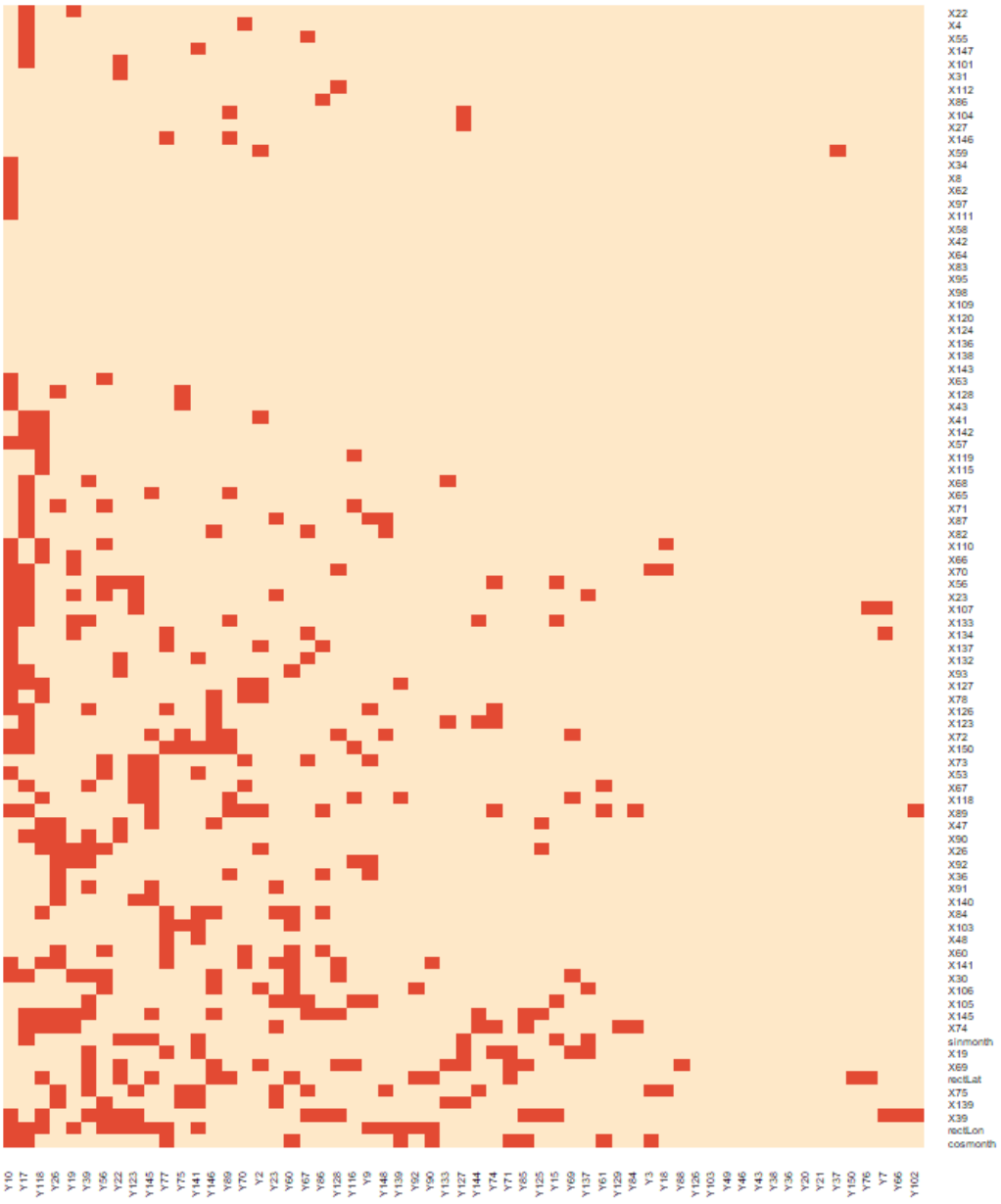

FIGURE A.10 – Variables sélectionnées (en rouge) pour l'étape R en choisissant les coefficients de régularisation avec le critère du minimum.

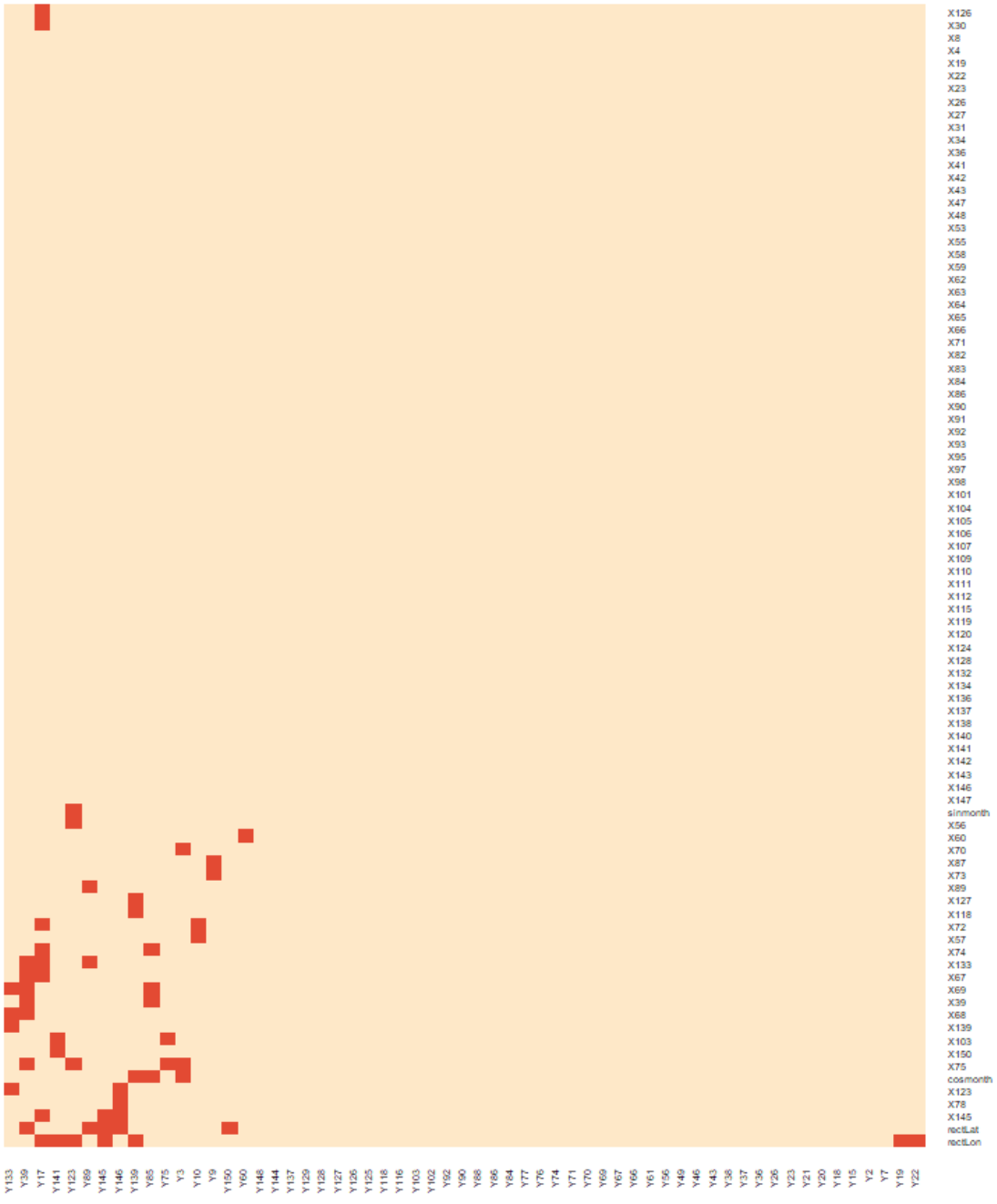

FIGURE A.11 – Variables sélectionnées (en rouge) pour l'étape R en choisissant les coefficients de régularisation avec le critère du "one-standard-error".

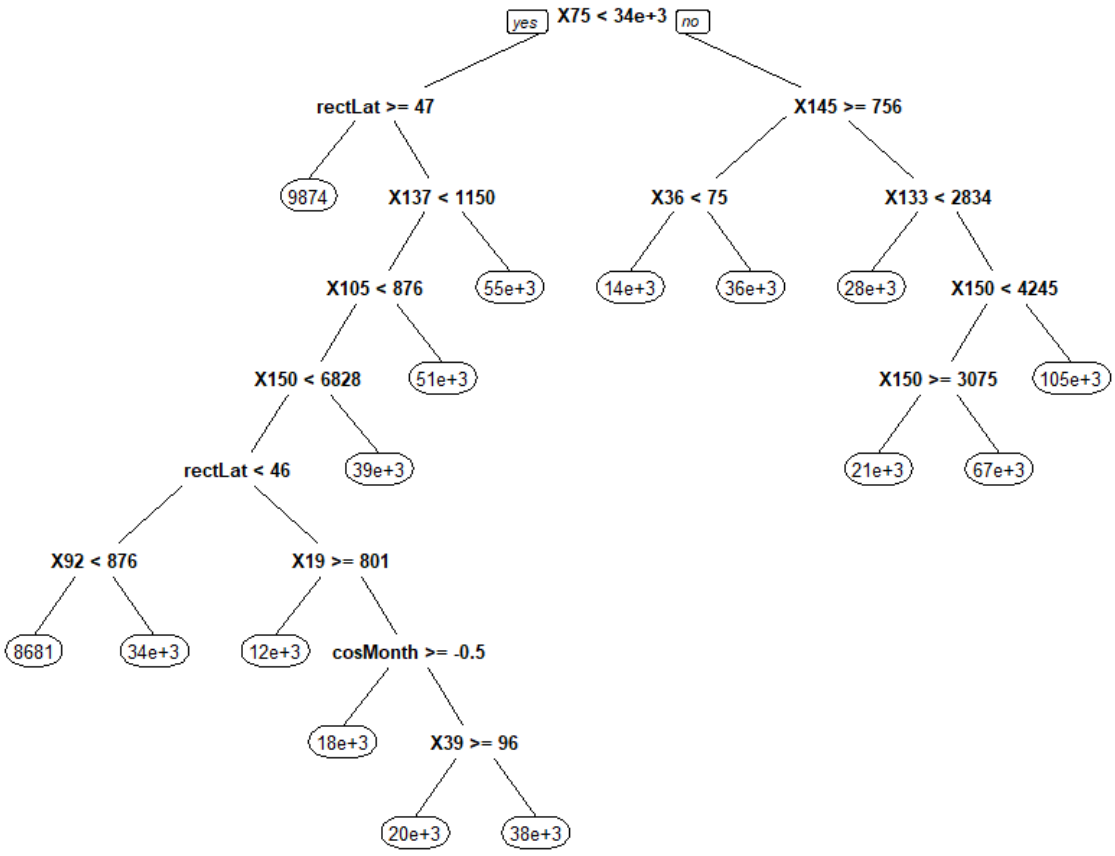

FIGURE A.12 – Arbre de régression pour les données de rejet du merlu commun (Y_{75}). L'arbre a été construit jusqu'au bout, avec la contrainte d'un minimum de 10 observations par feuilles. On a réduit cet arbre par *pruning* grâce à la validation croisée dans la figure 4.1.

Annexe B

Espèces rares

Espèce	Y_{pres}	$Y_{\text{moy rejet}}$	Y_{max}	X_{pres}	$X_{\text{moy debarq}}$	X_{max}
Acantholabrus palloni	1	414	414	0	0	0
Alosa alosa	4	3772	8375	1	326	326
Alosa fallax	4	5812	12480	0	0	0
Argentina silus	3	883	1200	0	0	0
Argyrosomus regius	2	953	1740	2	1355	1385
Arnoglossus rueppelii	2	1039	2000	0	0	0
Arnoglossus thori	1	25	25	0	0	0
Boops boops	1	520	520	0	0	0
Buenia jeffreysii	3	15	24	0	0	0
Callionymus	2	5750	10000	0	0	0
Ciliata mustela	1	1506	1506	0	0	0
Clupea harengus	4	1245	2233	0	0	0
Crangon crangon	2	959	998	1	21000	21000
Ctenolabrus rupestris	2	745	1200	0	0	0
Deltentosteus quadrimaculatus	1	1	1	0	0	0
Dicentrarchus labrax	0	0	0	16	2725	10910
Dicologlossa cuneata	5	470	904	3	280	736
Dipturus batis	1	2300	2300	0	0	0
Echiichthys vipera	2	2562	5027	0	0	0
Echinus esculentus	0	0	0	1	450	450
Eledone	1	4686	4686	0	0	0
Gadiculus argenteus	6	4809	16888	0	0	0
Gadus morhua	1	1776	1776	3	3845	10984
Gaidropsarus	2	825	1001	1	250	250
Galathea	4	3580	6873	0	0	0
Galathea squamifera	3	31062	76588	0	0	0
Galatheidae	0	0	0	5	7497	21000
Glyptocephalus cynoglossus	0	0	0	6	360	500
Goneplax rhomboides	1	344	344	0	0	0
Harengula clupeola	1	227	227	0	0	0
Helicolenus dactylopterus	4	590	1008	0	0	0
Homarus gammarus	0	0	0	9	945	2121
Illex	1	2600	2600	7	969	2800
Labrus bergylta	0	0	0	1	400	400

Lepidorhombus boscii	1	288	288	4	331	550
Leucoraja fullonica	0	0	0	2	903	1462
Leucoraja naevus	0	0	0	14	1802	7937
Liocarcinus holsatus	5	10934	17182	1	25200	25200
Loligo	0	0	0	3	4233	12000
Lophius	0	0	0	17	6374	16000
Maja squinado	0	0	0	2	1155	1260
Melanogrammus aeglefinus	3	1423	3582	18	2477	13242
Microstomus kitt	3	1157	1688	43	648	2095
Mola mola	1	5000	5000	0	0	0
Molva dypterygia	2	1066	1085	0	0	0
Molva macrophthalma	2	1209	2269	0	0	0
Molva molva	5	901	1668	18	5354	22494
Mullus barbatus barbatus	0	0	0	1	2080	2080
Mustelus asterias	2	1551	2382	73	7948	92560
Octopus	0	0	0	7	2092	5000
Pagellus	0	0	0	1	250	250
Pagellus acarne	1	979	979	0	0	0
Pagellus erythrinus	1	416	416	1	300	300
Pagrus pagrus	1	208	208	0	0	0
Palinurus elephas	2	444	488	4	552	762
Palinurus mauritanicus	0	0	0	1	368	368
Pecten maximus	3	1772	3197	0	0	0
Pegusa lascaris	4	2063	6336	0	0	0
Penaeus japonicus	0	0	0	1	50	50
Platichthys flesus	0	0	0	1	850	850
Pleuronectes platessa	1	3140	3140	28	862	2240
Pollachius pollachius	1	358	358	40	4789	15380
Pollachius virens	0	0	0	5	1189	3478
Pomatoschistus minutus	1	127	127	0	0	0
Raja brachyura	0	0	0	1	2896	2896
Raja clavata	1	242	242	9	3825	9653
Raja microocellata	0	0	0	1	3343	3343
Raja montagui	0	0	0	4	2276	3404
Raja undulata	1	84512	84512	0	0	0
Raniceps raninus	1	963	963	0	0	0
Scomber colias	2	235	296	0	0	0
Scophthalmus maximus	0	0	0	6	2022	3346
Scophthalmus rhombus	0	0	0	4	1103	1319
Scorpaena notata	3	18000	36000	0	0	0
Scorpaena scrofa	3	2406	4000	0	0	0
Scyliorhinus stellaris	2	22083	40818	1	520	520
Sepiola atlantica	1	300	300	0	0	0
Sepiola robusta	1	862	862	0	0	0
Solea senegalensis	0	0	0	2	425	580
Spondyliosoma cantharus	0	0	0	15	712	1632
Squalus acanthias	0	0	0	1	7000	7000
Torpedo marmorata	4	4249	8700	1	800	800
Trachurus mediterraneus	0	0	0	2	1414	1446

Trigla lyra	0	0	0	1	320	320
Trigloporus lastoviza	0	0	0	2	278	348
Trisopterus esmarkii	7	2006	4800	0	0	0
Umbrina canariensis	0	0	0	3	282	300
Zeugopterus regius	1	336	336	0	0	0

TABLE B.1: Informations sur la distribution des espèces rares non traitées par les modèles (moins de 5 présences dans l'ensemble d'entraînement). *pres* désigne le nombre de présences, *moy* la moyenne des quantités > 0 et *max* la quantité maximale présente.