

Le modèle linéaire

Application à la modélisation de la distribution des espèces marines

1 Question écologique et cas d'application

1.1 Distribution des espèces et niche écologique

La distribution des espèces varie en fonction des conditions environnementales (facteurs abiotiques - température, salinité ou profondeur) et de processus d'interaction entre individus et espèces (facteurs biotiques - compétition, prédation ou coopération). Comprendre ces mécanismes est essentiel pour relier la présence ou l'abondance d'une espèce aux caractéristiques de son habitat. En général, les modèles statistiques utilisés pour décrire la distribution des espèces permettent d'inférer la relation d'une espèce à son environnement et se concentrent sur les facteurs abiotiques de la distribution des espèces. C'est ce qui est généralement appelé **niche écologique**. Dans le cadre de ce cours, nous cherchons à relier la distribution des poissons du Golfe de Gascogne aux covariables environnementales influençant la répartition de ces espèces (*i.e.*, inférer la niche écologique des espèces). Une méthode standard pour inférer la niche écologique d'une espèce est le **modèle linéaire**.

1.2 Données d'application

L'étude de la niche écologique des espèces marines repose en grande partie sur les données issues de campagnes océanographiques. Ces campagnes permettent de récolter des données d'abondance et de biomasse des espèces sur l'ensemble d'un écosystème donnée - par exemple Golfe de Gascogne (GdG) / Mer Celtique (MC). Ces données sont cruciales pour le suivi des espèces marines afin d'évaluer le bon état écologique des populations exploitées. Elles permettent notamment de mettre en relation la biomasse ou l'abondance des espèces et les covariables environnementales influençant la répartition des ces espèces.

Pour ce cours, nous allons étudier les données issues de la campagne **EVHOE**. Les données EVHOE (Evaluation Halieutique Ouest de l'Europe) sont des données échantillonnées chaque année en Octobre/Novembre. Cette campagne cible les espèces benthico-démersales du golfe de Gascogne (GdG) et de Mer Celtique (MC). L'échantillonnage de la zone d'étude est stratifié

suivant les classes de profondeur et les grandes unités écologiques du GdG et de MC. La description des données est donnée en annexe.

2 Le modèle linéaire

Dans un premier temps, on introduit le modèle linéaire qui va permettre de décrire la niche écologique des espèces à partir des données EVHOE. On considère que la distribution des espèces dépend de la profondeur et de la zone géographique (Mer Celtique Nord/Centre et Sud, Golfe de Gascogne Nord et Sud). En particulier, la profondeur a un effet quadratique sur la quantité de poisson capturée et il y a quatre niveaux de facteurs pour la zone géographique (**StrataCcn** pour la Mer Celtique centre et nord, **StrataCs** pour la Mer Celtrique sud, **StrataGn** pour le Golfe de Gascogne nord, **StrataGs** pour le Golfe de Gascogne sud).

Le modèle linéaire correspondant s'écrit:

$$y_{ij} = \alpha + \beta_1 \cdot d_i + \beta_2 \cdot d_i^2 + \gamma_j + \epsilon_{ij}$$

où :

- y_{ij} est le *log* de la biomasse capturée au point i pour le niveau de facteur j ,
- d_i est la profondeur,
- α est l'intercept,
- β_1 et β_2 sont les paramètres associés à la profondeur,
- $(\gamma_j)_{j \in \{1, \dots, 4\}}$ sont les paramètres associés à la localisation en MC/GdG. Pour un soucis d'identifiabilité, γ_1 est fixé tel que $\gamma_1 = 0$. L'effet de la strate **StrataCcn** est capturée par l'intercept α .
- ϵ_{ij} est une variable aléatoire gaussienne de variance σ^2 tel que:

$$\epsilon_{ij} \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

Sous forme matricielle, le modèle peut s'écrire:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{E}$$

où:

$$\mathbf{Y} = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n4} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{n4} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & d_1 & d_1^2 & (\delta_1)_1 & (\delta_2)_1 & (\delta_3)_1 & (\delta_4)_1 \\ 1 & d_2 & d_2^2 & (\delta_1)_2 & (\delta_2)_2 & (\delta_3)_2 & (\delta_4)_2 \\ 1 & d_3 & d_3^2 & (\delta_1)_3 & (\delta_2)_2 & (\delta_3)_3 & (\delta_4)_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & d_i & d_i^2 & (\delta_1)_i & (\delta_2)_i & (\delta_3)_i & (\delta_4)_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & d_n & d_n^2 & (\delta_1)_n & (\delta_2)_n & (\delta_3)_n & (\delta_4)_n \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{D}^{(1)} & (\mathbf{D}^2)^{(1)} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{D}^{(2)} & (\mathbf{D}^2)^{(2)} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{D}^{(3)} & (\mathbf{D}^2)^{(3)} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \mathbf{0}_{n_3} \\ \mathbf{1}_{n_4} & \mathbf{D}^{(4)} & (\mathbf{D}^2)^{(4)} & \mathbf{0}_{n_4} & \mathbf{0}_{n_4} & \mathbf{0}_{n_4} & \mathbf{1}_{n_4} \end{bmatrix}$$

- $\mathbf{1}_n$: vecteur de uns de taille n . $\mathbf{0}_n$, vecteur de zéros de taille n .
- $(d_1, \dots, d_n)^\top$: profondeur de chaque point d'échantillonnage, associée au paramètre β_1 .
- $(d_1^2, \dots, d_n^2)^\top$: profondeur au carré, associée au paramètre β_2 .
- $((\delta_1)_1, \dots, (\delta_1)_n)^\top$: indicatrice pour la première zone géographique ‘StrataCcn’, associée à γ_1 . Il peut être fixé à 0 pour assurer l'identifiabilité du modèle. Le nombre d'échantillon associé à ce niveau de facteur est noté n_1 . Les mêmes notations s'appliquent pour $(\delta_2)_i$, $(\delta_3)_i$, $(\delta_4)_i$.
- $\mathbf{D}^{(1)}$: vecteur de l'ensemble des profondeurs pour la première zone géographique. $(\mathbf{D}^2)^{(1)}$: vecteur de l'ensemble des profondeurs au carré pour la première zone géographique. Les mêmes notations s'appliquent pour les autres zones géographiques.

Dans la suite, y_i est la réalisation de cette variable aléatoire. \mathbf{x}_i désigne la i^e ligne de la matrice \mathbf{X} et ϵ_i le i^e terme du vecteur \mathbf{E} .

Le modèle linéaire appliqué sur les principales espèces capturées par la campagne EVHOE donne les estimations suivantes.

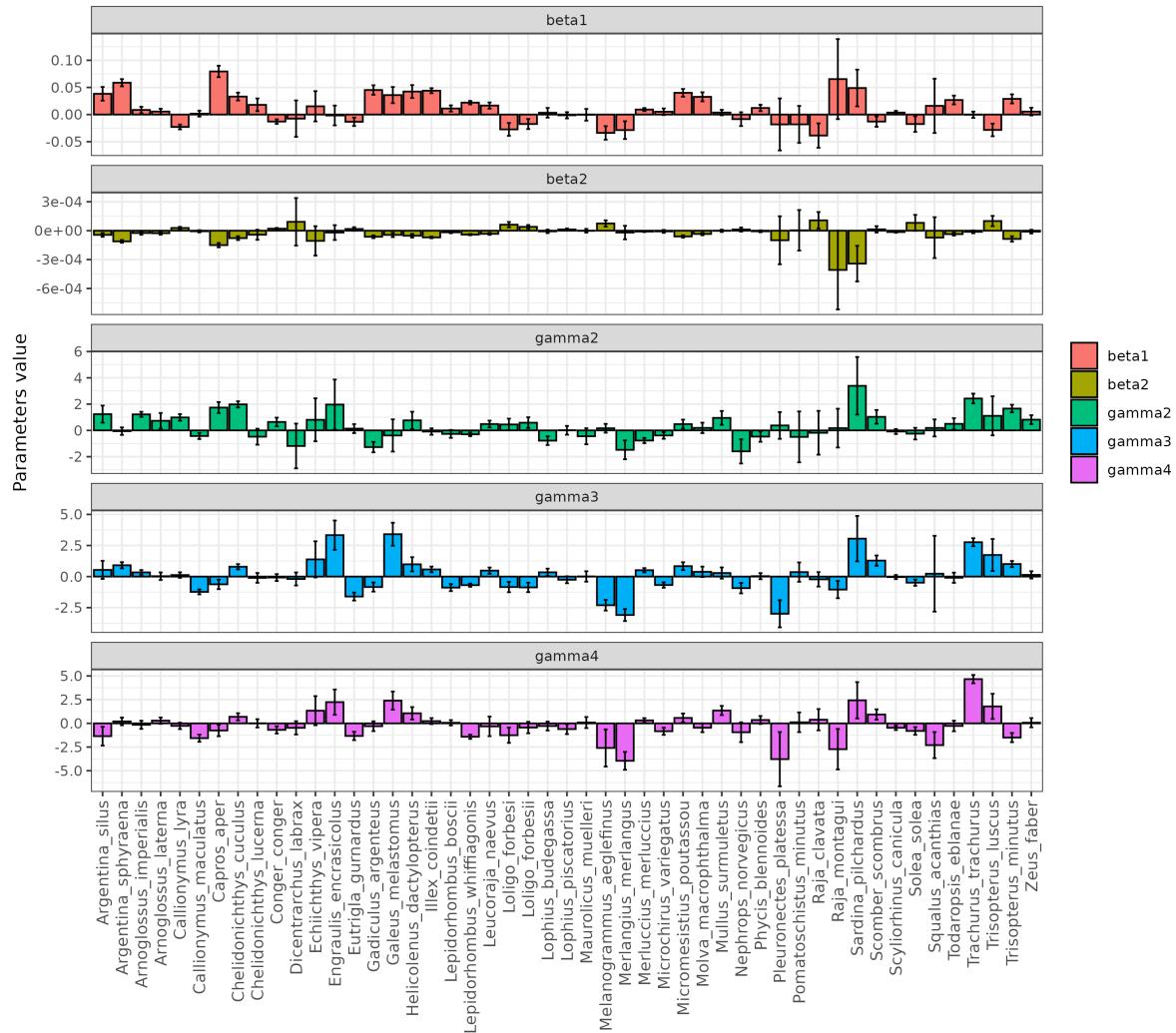


Figure 1: Paramètres du modèles linéaire pour les principales espèces observées de la campagne EVHOE.

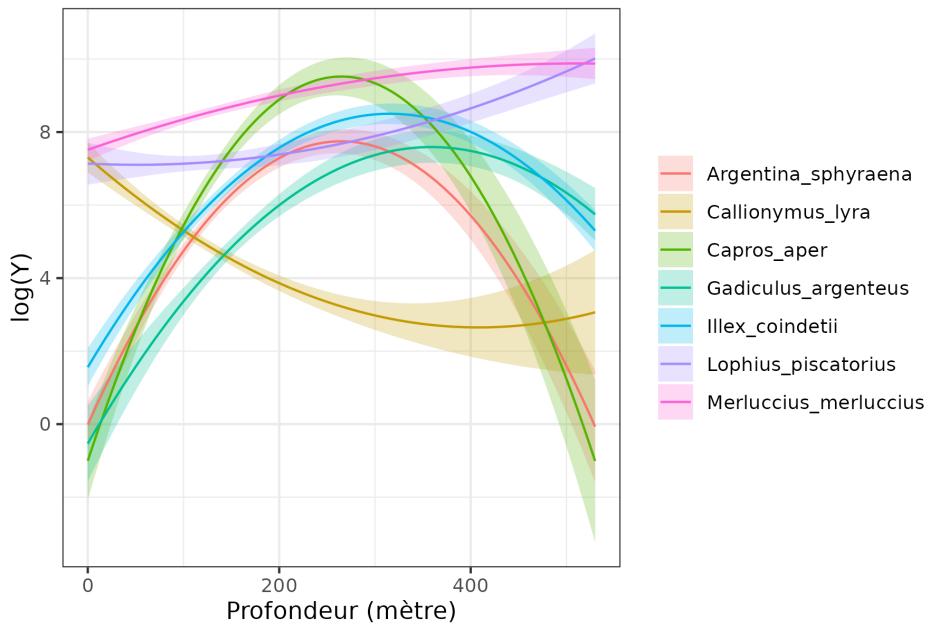


Figure 2: Réponse de différentes espèces à la profondeur.

Sources bibliographiques

Bel, L., Daudin, J. J., Etienne, M., Lebarbier, E., Mary-Huard, T., Robin, S., & Vuillet, C. (2016). Le Modèle Linéaire et ses Extensions. [Lien](#).

3 Annexes

3.1 Description des données

Les poissons sont échantillonnées à l'aide d'un chalut ; ils sont comptés, pesés, sexés pour tout ou partie du trait de chalut. Les données entre 2008 et 2019 sont stockés dans le fichier `EVHOE_2008_2019.RData`. Il est constitué de trois data frame:

- `Save_Datras$datras_HH.full` regroupe les principales informations de chaque trait de chalut (e.g. localisation, période de relevé)
 - Year: année
 - long: longitude
 - lati: latitude
 - StNo: numéro de station
 - HaulNo: numéro du trait de chalut
 - Depth: profondeur
 - Distance: distance parcourue pour un trait de chalut (en mètres). Il y a des NA dans cette colonne (données manquantes). Dans ce cas, on prend la moyenne de la distance des autres traits de chaluts pour remplacer les NA.
- `Save_Datras$datras_sp.HL.full` regroupe le poids et les abondances sur l'ensemble d'un trait de chalut de chaque combinaison ‘trait de chalut x espèce x classe de taille x sexe’ (données ré-haussées)
 - Year: année
 - long: longitude
 - lati: latitude
 - StNo: numéro de station
 - HaulNo: numéro du trait de chalut
 - scientificname: nom scientifique
 - LngtClass: classe de taille
 - TotalNo: comptages (nombre d'individus par combinaison de facteur)
- `Save_Datras$datras_sp.CA.full` regroupe les données de mesures individuelles d'un sous-échantillon du trait de chalut. Une ligne correspond à un individu. Ces données regroupent les données individuelles de taille, de poids, de sexe. Nous n'utiliserons pas ces données dans ce projet.



Figure 3: Récolte des données EVHOE.

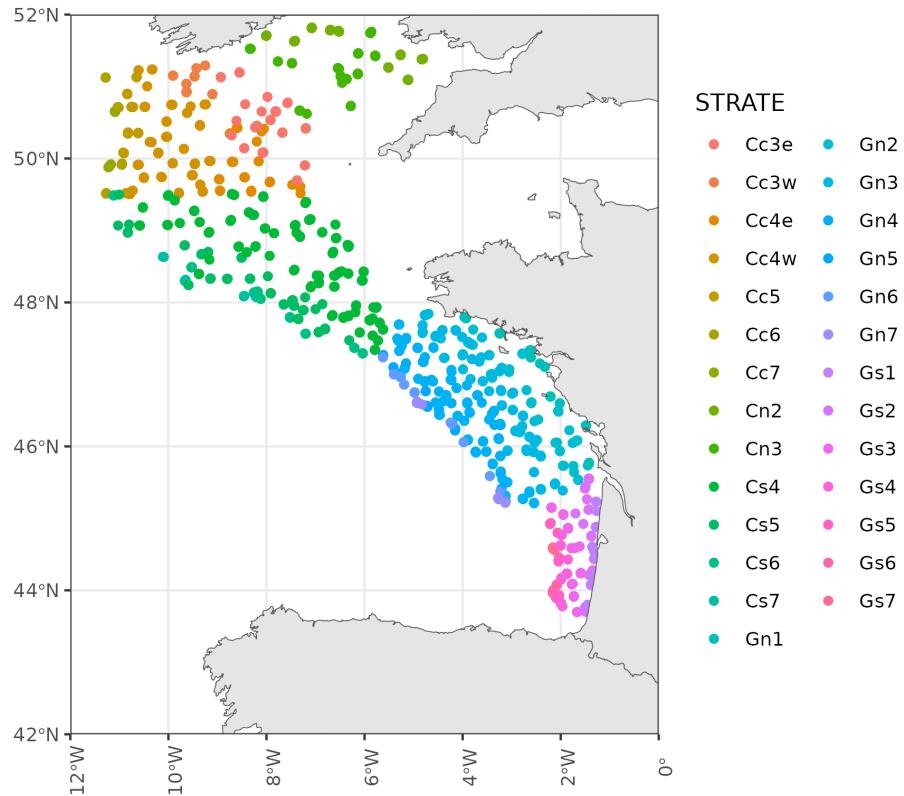


Figure 4: Points échantillonnés par la campagne EVHOE toutes années confondues.

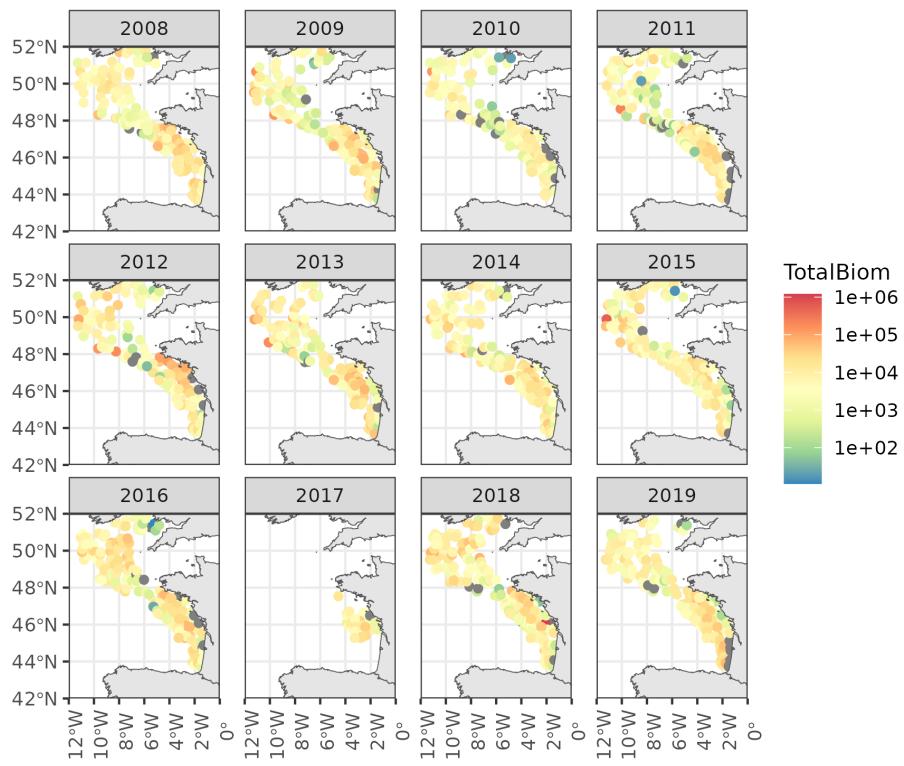


Figure 5: Données de biomasse pour le merlu (*Merluccius merluccius*) en kg