Formation à R

Modélisation avec les GLM

Sébastien Rochette, ThinkR

ThinkR

Table des matières

1. Préface

La version d'origine de cette formation a été créée par Olivier Le Pape et Étienne Rivot à Agrocampus Ouest (Rennes, France). Depuis mon doctorat dans leur équipe, je mets à jour constamment cette formation au gré de ma recherche et de l'évolution du logiciel R.

```
# Generated with R and rmarkdown: Roadmap version - Students
```

2. Présentation de l'étude

Le contexte et les objectifs de votre étude définissent le type de modélisation que vous allez mettre en place sur votre jeu de données.
Ici, nous utilisons les modèles linéaires généralisés pour produire une carte de distribution moyenne de la nourricerie de soles communes de la baie de Vilaine.

2.1. Contexte

- Les zones côtières et les estuaires sont des habitats halieutiques essentiels
  - Zones à forte production
  - Nourriceries
  - Zones restreintes avec de fortes densités (Fig. 1)
- Pression anthropique élevée
  - Perte de surface disponibles (Fig. 2a)

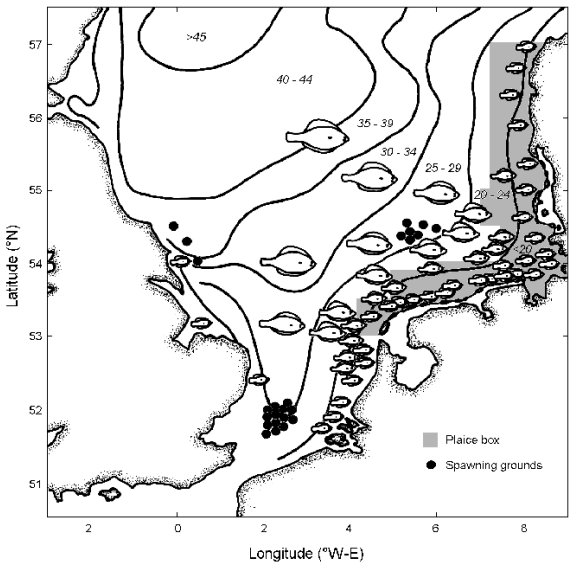

Figure 1 – Plaice box (Rijnsdorp *et al.*)

- Qualité des habitats alterée (Fig. 2b)

- Impact sur le renouvellement des populations
  - Jeune stades = Gouleau d'étranglement
  - La taille et la qualité des nourriceries côtières influent sur la production de juvéniles

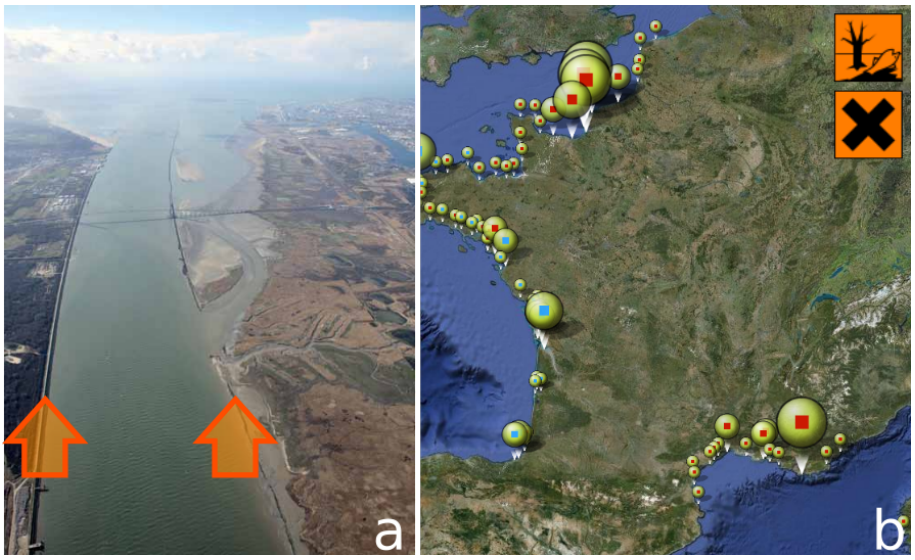

Figure 2 – (a) L'estuaire de la Seine. (b) Niveau de contamination chimique le long des côtes françaises (Ifremer, 2011)

2.2. Objectifs

Déterminer les facteurs ayant une influence sur la distribution des poissons plats (*Solea solea*) en Baie de Vilaine et cartographier la distribution moyenne des densités.

- Cartographier les habitats potentiels nécessite:
  - Connaissance des habitats de juvéniles
  - Campagnes d'échantillonnage dans la zone d'étude
  - Connaissance des covariables environnementales ayant potentiellement de l'influence
    - Cartes exhaustives des covariables environnementales
- Une approche statistique en deux étapes

- Modèle statistique reliant les densités aux covariables
- Prédire les habitats potentiels

2.3. Données

Campagne standardisée de chalut à perche dans la baie de Vilaine (Fig. 3)

- 1984 – 2010
- En autumne
- Juvéniles de l'année (Âge 0)
  - Nb individus / 1000m^2

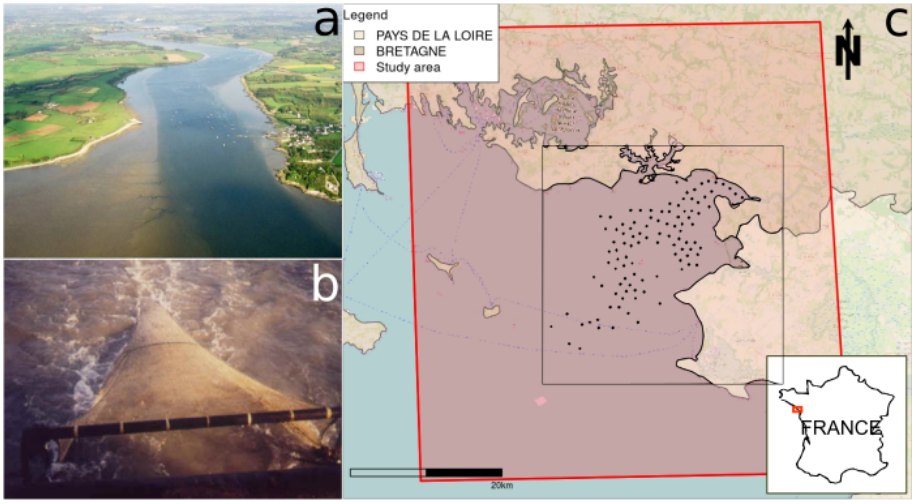

Figure 3 – (a) L'estuaire de la Vilaine. (b) Chalut à perche. (c) Situation des stations d'échantillonnage.

2.4. Covariables

- Bathymétrie (Fig. 4a)
  - MNT à 1000m de résolution
  - Projection Mercator
- Structure sédimentainre (Fig. 4b)
  - Fichier shape de polygones
  - Coordonnées géographiques
- Zones biologiques (Fig. 4c)
  - Combinaison bathymétrie, sédiment, habitat
  - Fichier shape de polygones
  - Coordonnées géographiques

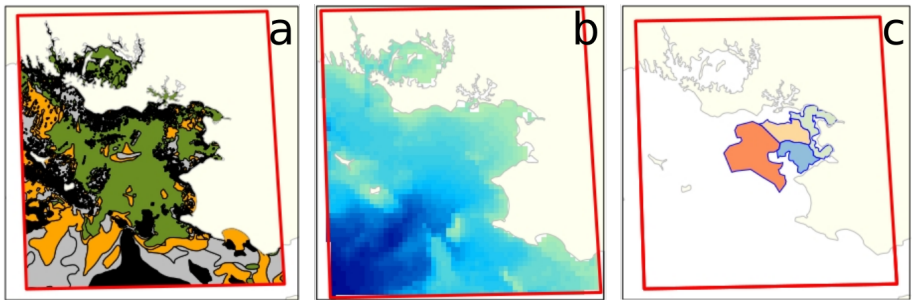

Figure 4 – Covariables en baie de Vilaine. (a) Structure sédimentaire, (b) Bathymétrie et (c) Zones biologiques.

2.5. Ajuster un modèle de distribution d'espèces

- Croiser les données avec les cartes de covariables
  - Utiliser un modèle linéaire
- Utiliser les cartes des covariables pour la prédiction (Fig. 5)
  - Une prédiction pour chaque cellule d'un raster

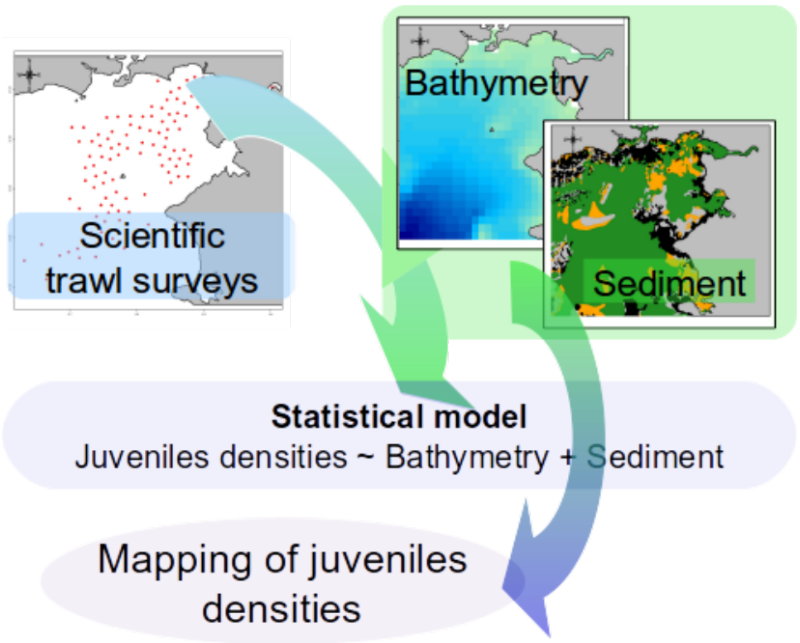

Figure 5 – Procédure pour un modèle de distribution d'espèce

2.6. Exploration des données

Prenez le temps d'explorer vos données avant toutes analyses

- Explorer les données et les covariables
  - Explorer le plan d'échantillonnage
  - Explorer les liens potentiels entre les densités et les covariables
  - Explorer les futurs paramètres de modélisation (interactions, distributions)

Souvenez-vous toujours des objectifs de votre étude !

Question : Que recherchons-nous dans cette exploration ?

3. Préparation

3.1. Structure des dossiers

Il convient de toujours conserver les fichiers originaux : les reprojections entraînent toujours quelques pertes, mieux vaut revenir aux originaux lorsque c'est possible.

L'arborescence de votre dossier de travail est la suivante :

- 01_Original_data
  - DEPARTEMENTS
  - Sedim_GDG_wgs84

  - bathy_GDG_1000_merc (and co)
  - Data_Vilaine_solea.csv
- 02_Outputs
- 03_Figures
- 04_Functions

3.2. Débutons avec R

- Créer un projet Rstudio dans le dossier principal de travail.
- Ouvrez le script R : "Classic_PresAbs_Positive_HSI_Student.R"
- Lister les différents sous-dossier de travail au début de votre script R

```
# Define working directories ----------------------------------------------------
WD <- here()
# Folder of original files
origWD <- here("01_Original_data")
# Folder for outputs
saveWD <- here("02_Outputs")
# Folder where to save outputs from R
figWD <- here("03_Figures")
# Folder where complementary functions are stored
funcWD <- here("04_Functions")
```

4. Modèle Delta

4.1. Étapes

Le modèle sur les données complètes n'était pas satisfaisant. Pour mieux prendre en compte (1) les données d'absences et (2) les fortes valeurs de densités, nous allons utiliser un approche Delta. Le modèle Delta sépare les données en deux sous-groupes, un pour la présence-absence, l'autre pour les densités lorsqu'il y a présence.

- Construction d'un modèle de présence / absence
  - Distribution binomiale
  - Prédiction de probabilités de présence
- Construction d'un modèle sur données positives
  - Distribution à définir
  - Prédiction des densités lorsqu'il y a présence

Puisque les modèles sont ajustés séparément, ils peuvent inclure des covariables différentes.

- L'approche Delta couple les deux sous-modèles
  - Sous-modèle Binomial : $p_{0/1}$
  - Sous-modèle positif : $Dens_{+}$
  - Couplage: $Density = p_{0/1} \cdot Dens_{+}$

4.2. Sous-modèle sur données positives

4.2.1 Étapes

La procédure à adopter avec le sous-groupe de données est la même qu'avec le jeu de données complet.

- Créer un sous-jeu de données contenant uniquement les observations positives
- Explorer ce nouveau jeu de données

  - Explorer les effets potentiels des covariables
  - Explorer les potentielles loi de distributions
  - Tester les interactions
- Choisir le meilleur modèle

.

4.2.2 Validation du modèle

En utilisant les différents indices présenté précédemment, vous choisissez le modèle qui s'ajuste le mieux à vos données. C'est donc le meilleur modèle pour décrire vos observations. Dans notre cas, nous souhaitons aussi utiliser ce modèle pour faire de la prédiction, ce qui nécessite de sélectionner un modèle qui donne de bonnes prédictions sur des données non-utilisées pour l'ajustement du modèle. Pour cela, nous pouvons utiliser la validation croisée :

- Ajuster un modèle sur 90% des données par exemple
- Utiliser le modèle ajusté pour faire une prédiction pour les 10% restants
- Comparer les prédictions aux observations
- Choisir le modèle ayant le meilleur indice de comparaison

Vous pourriez utiliser le coefficient de corrélation entre les prédictions et les données de validation comme un indice de qualité d'ajustement pour sélectionner le meilleur modèle. Cependant, l'erreur quadratique moyenne (MSE = Mean Squared Error) voire sa racine (RMSE = Root MSE) est l'indice recommandé. Il mesure la distance moyenne d'une observation à sa prédiction.
- *Le modèle sélectionné sur la base de l'AIC est-il toujours le meilleur modèle avec le RMSE ?*

4.3. Sous-modèle Binomial

4.3.1 Étapes

La procédure à adopter avec le sous-groupe de données est la même qu'avec le jeu de données complet.

- Créer les observations de présence-absences à partir du jeu de données
- Explorer ce nouveau jeu de données
- Utiliser une distribution binomiale
  - Tester les covariables, les interactions, les fonctions de lien, les critères de qualité
- Choisir le meilleur modèle

4.3.2 Exploration

4.3.3 Ajuster un modèle binomial avec une fonction de lien

Le choix de la distribution pour un modèle de présence-absence est simple, c'est un modèle binomial. Cepedant, un modèle est généralement ajusté sur la base de résidus Gaussiens. Pour ajuster un modèle binomial, les données doivent être transformées de telle sorte qu'on puisse ajuster un modèle linéaire Gaussien classique dessus. Pour cela, nous utilisons une fonction de lien. La fonction de lien classique d'un modèle binomial est la fonction logit, mais ce n'est pas la seule. Vous pouvez tester cloglog, probit ou cauchit.

4.3.4 Qualité d'ajustement d'un modèle binomial

Une mesure couramment utilisée pour la qualité d'ajustement d'un modèle binomial est "l'aire sous la courbe" (AUC : Area Under the Curve). Un objectif des modèles binomiaux étant de prédire un succès ou un échec, et non pas seulement une probabilité de succès, on peut vouloir définir un seuil (intuitivement 0.5 par exemple) qui transforme la probabilité de présence en présence ou absence. L'AUC est en quelque sorte une probabilité de classer correctement les présences et absences. Une définition plus complète serait :

La probabilité moyenne pour qu'une observation=1 et une observation=0 choisies de manière aléatoire dans le jeu de données montrent une probabilité de présence prédite supérieure pour l'observation=1 par rapport à celle de l'observation=0

Ainsi, $AUC = 1$ montrerait un modèle "parfait", mais $AUC = 0.5$ montrerait un modèle plus mauvais que le hasard.

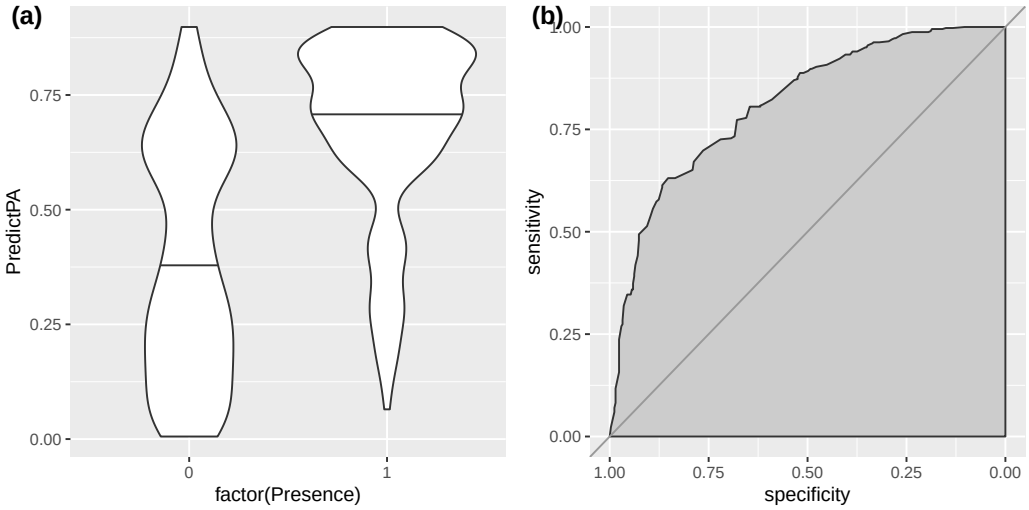

Figure 6 – (a) Prédiction vs Observations. (b) Courbe ROC d'un modèle binomial

4.3.5 Choix du meilleur seuil

- *Le modèle sélectionné sur la base de l'AIC est-il toujours le meilleur modèle avec l'AUC sur les données de validation ?*

4.4. Couplage des deux sous-modèles

4.4.1 L'approche Delta

L'approche Delta est la méthode pour coupler les deux sous-modèles. En réalité, les deux modèles sont simplement multipliés l'un à l'autre.

Le couplage des deux sous-modèles c'est :

- Sous-modèle binomial: $p_{0/1} \sim Bathymetry + Sediment$
- Sous-modèle positif: $Dens_+ \sim Bathymetry$
  - Si log-transformation: $Dens_+ = exp(log(Y_+)) \times exp(-0.5 \cdot \sigma^2 \cdot log(Y_+))$
- Couplage: $Density = p_{0/1} \cdot Dens_+$

5. Modèle d'habitat

5.1. Étapes

La réalisation d'une carte de distribution d'espèce (Fig. 7) nécessite :

- Un modèle d'habitat potentiel
  - Indice de qualité d'habitat
  - Modèle: $Density \sim Bathymetry + Sediment$
- Les cartes complètes des covariables
- Une carte des prédictions du modèle

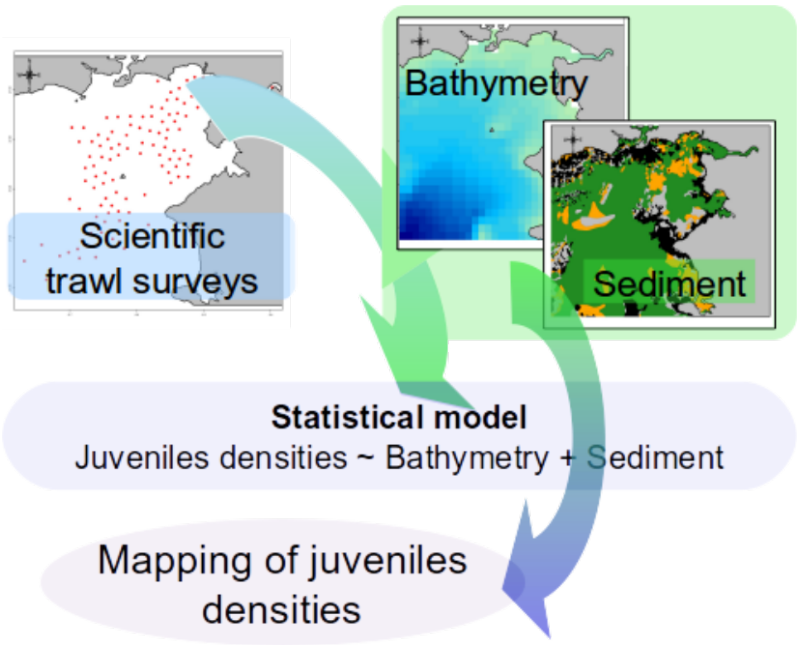

Figure 7 – Procédure pour cartographier une distribution d'espèce

Une manière simple de réaliser la carte des prédictions est de créer un raster qui rassemble l'information de toutes les cartes des covariables nécessaires, puis d'utiliser le modèle pour prédire dans chaque cellule du raster (Fig. 8).

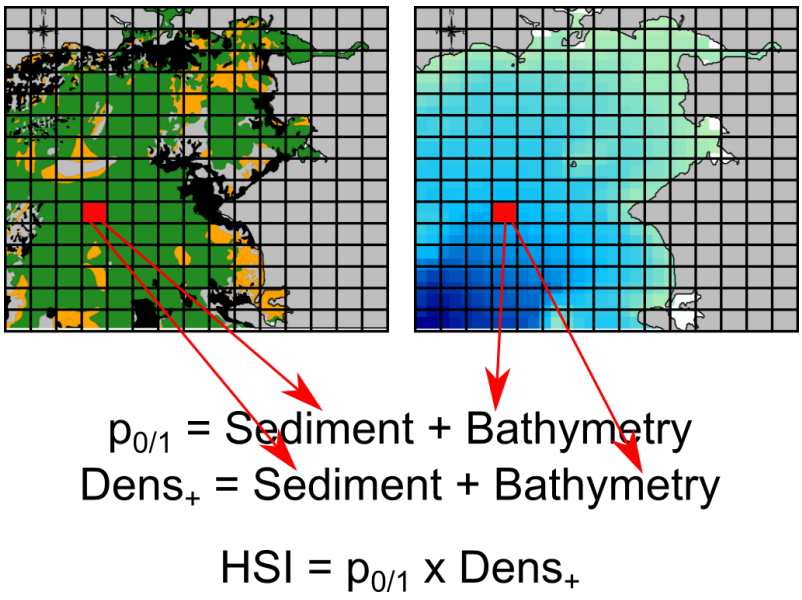

Figure 8 – Prédiction de densité pour chaque cellule d'une carte au format raster

5.2. Préparation des données

Pour pouvoir faire les prédictions dans le raster, ses couches doivent avoir le même nom que les colonnes du jeu de données. De plus, un raster est une matrice de valeurs numériques, ce qui oblige à convertir les covariables en classe en valeurs numériques, de telle sorte que les niveau de facteur du raster correspondent à ceux des données, et donc existent dans les modèles. *Soyez prudents avec la conversion vers des valeurs numériques, les données doivent rester au format facteur et ne doivent pas être utilisées comme valeurs numériques dans les modèles !*

5.3. Prédictions

La fonction `predict` a une méthode pour pouvoir être utilisée directement sur un objet Raster (Fig. 9).

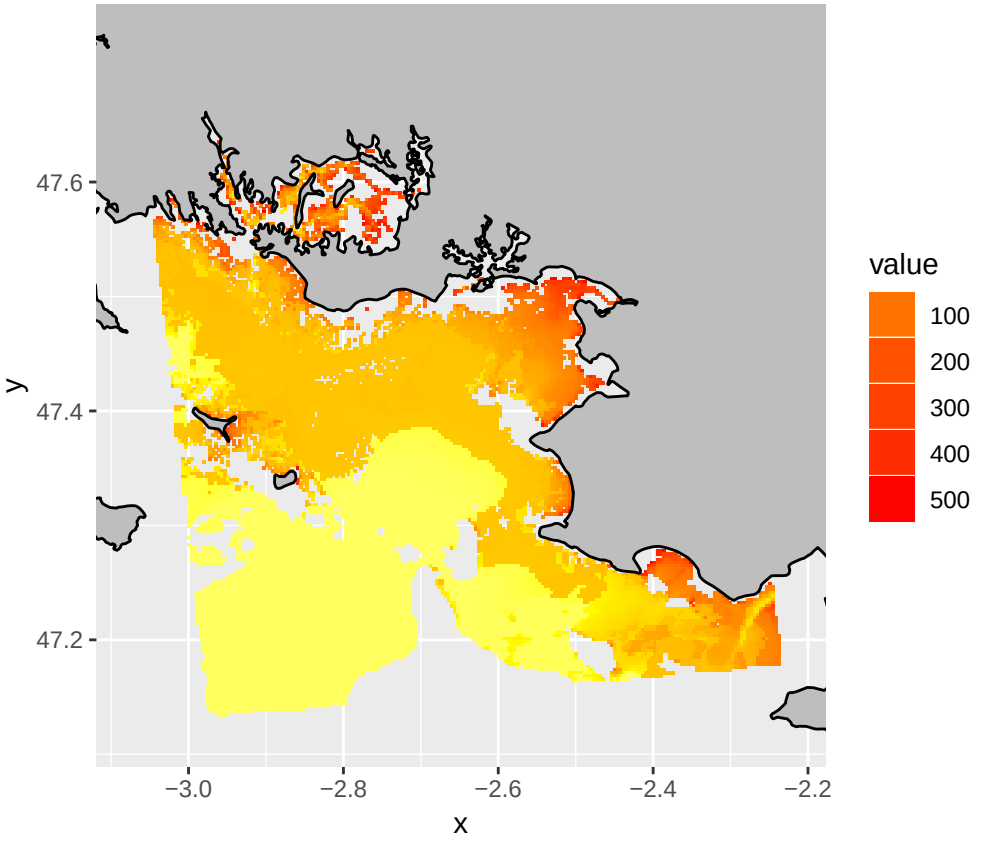

Figure 9 – Prédiction des densités de soles en baie de Vilaine. Échelle de couleur en fonction des quantiles des données originales (10%, 50%, 75%, 95%).

6. Conclusion

6.1. Modélisation

- **Importance de l'exploration des données**
  - Validation des données
  - Loi de distribution
  - Options pour les modèles
- **Étude des modèles : une approche itérative**
  - Choix de la loi de distribution
  - Choix des combinaisons de covariables
  - Vérification des hypothèses
    - Analyse des résidus
    - Critères de qualité d'ajustement
- **Gardez toujours vos objectifs en tête !**

6.2. Modèle Delta

- **Utilité d'un modèle Delta**
  - Les données brutes ne peuvent être modélisées
  - Sens biologique

  - Présence & densités: pas forcément les mêmes covariables
- **Utilisation d'un modèle Delta**
  - Prédictions utiles
  - Les paramètres des sous-modèles n'ont pas de sens dans le modèle couplé
- **Alternatives**
  - Autres modèles zero-inflated ? Distribution tweedie ?
  - GAM (Attention aux données nécessaires et à l'interprétation)
  - Régression quantile (habitat préférentiels)
  - Random forest (Attention à votre question)

6.3. Modèles de distribution d'espèces

- **Outils utiles pour les connaissances biologiques et pour la gestion**
  - Modèle Delta approprié pour les données d'espèces marines
  - Fiable si utilisé avec précaution
    - Ce ne sont que des corrélations...
- **Cette formation est un exemple simple**
  - D'autres perspectives
    - Ajouter une covariable biotique
    - Approche multi-spécifique
    - Pressions anthropiques ?
- **D'autres outils existent**
  - Votre question détermine l'outil à utiliser

6.4. Exemples d'applications

- Effet de la destruction d'habitat sur la biomasse de juvéniles (Fig. 10)
  - Cas de la sole commune dans l'estuaire de Seine
  - Comparaison entre 1850 et 2004
    - Perte en surface : 33%
    - Perte en biomasse : 42%

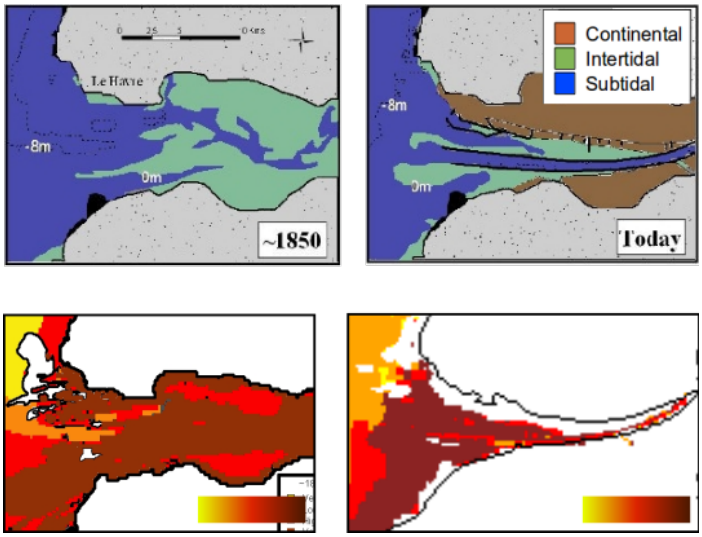

Figure 10 – Modélisation des effets de la destruction d'habitats sur la biomasse de sole en Seine. Rochette, S., Rivot, E., Morin, J., Mackinson, S., Riou, P., Le Pape, O. (2010). Effect of nursery habitat degradation on flatfish population renewal. Application to Solea solea in the Eastern Channel (Western Europe). Journal of sea Research, 64 : 34-44.

- Estimation de stock pour la gestion (Fig. 11)
  - Cas des laminaires du Parc marin d'Iroise
  - Estimation des biomasses
  - Validation avec les pêcheurs

- Proposition de gestion spatialisée

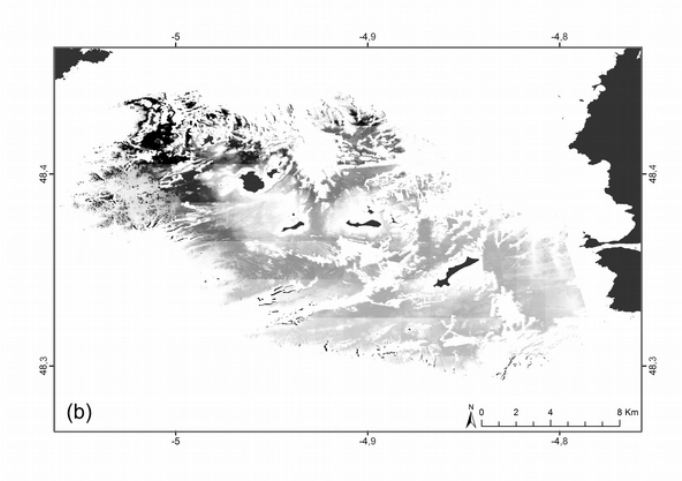

Figure 11 – Estimation spatialisée des biomasses de laminaires dans le parc marin d'Iroise pour la gestion de la ressource. Bajjouk T., Rochette S., Ehrhold A., Laurans M., Le Niliot P. (2015). Multi-approach mapping to help spatial planning and management of the kelp species L. digitata and L. hyporborea: Case study of the Molène archipelago, Brittany. Journal of Sea Research.