



Formation à R

Modélisation avec les GLM

SÉBASTIEN ROCHETTE, THINKR

ThinkR

Table des mati  res

1	Pr��face	1
2	Pr��sentation de l��tude	1
2.1	Contexte	1
2.2	Objectifs	2
2.3	Donn��es	3
2.4	Covariables	3
2.5	Ajuster un mod��le de distribution d��esp��ces	4
2.6	Exploration des donn��es	4
3	Pr��paration	4
3.1	Structure des dossiers	4
3.2	D��butons avec R	5
4	Mod��le Delta	5
4.1	��tapes	5
4.2	Sous-mod��le sur donn��es positives	5
4.3	Sous-mod��le Binomial	7
4.4	Couplage des deux sous-mod��les	11
4.5	Conclusion	12
5	Mod��le d��habitat	13
5.1	��tapes	13
5.2	Pr��paration des donn��es	14
5.3	Pr��dictions	14
6	Conclusion	15
6.1	Mod��lisation	15
6.2	Mod��le Delta	15
6.3	Mod��les de distribution d��esp��ces	16
6.4	Exemples d��applications	16

1. Pr  face

La version d  origine de cette formation a   t   cr   e par Olivier Le Pape et   tienne Rivot    Agrocampus Ouest (Rennes, France). Depuis mon doctorat dans leur   quipe, je mets    jour constamment cette formation au gr   de ma recherche et de l  volution du logiciel R.

Generated with R and rmarkdown: Roadmap version - Teacher

2. Pr  sentation de l  tude

Le contexte et les objectifs de votre   tude d  finissent le type de mod  lisation que vous allez mettre en place sur votre jeu de donn  es.

Ici, nous utilisons les mod  les lin  aires g  n  ralis  s pour produire une carte de distribution moyenne de la nourricerie de soles communes de la baie de Vilaine.

2.1. Contexte

- Les zones c  ti  res et les estuaires sont des habitats halieutiques essentiels
 - Zones    forte production
 - Nourriceries
 - Zones restreintes avec de fortes densit  s (Fig. 1)
- Pression anthropique   lev  e

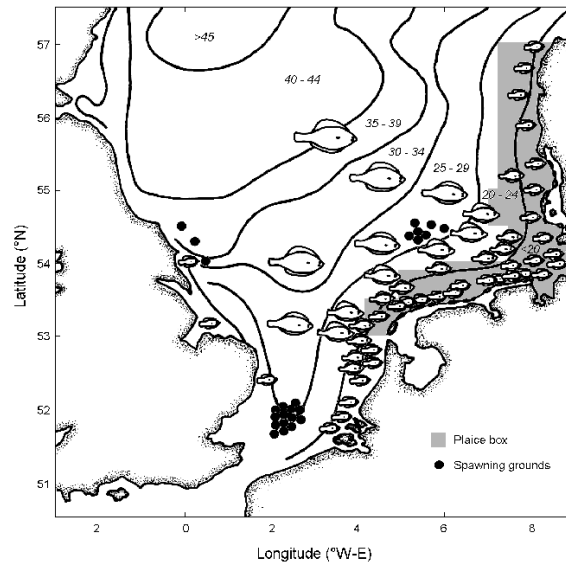


Figure 1 – Plaise box (Rijnsdorp *et al.*)

- Perte de surface disponibles (Fig. 2a)
- Qualité des habitats altérée (Fig. 2b)
- Impact sur le renouvellement des populations
 - Jeune stades = Goulet d'étranglement
 - La taille et la qualité des nourriceries côtières influent sur la production de juvéniles

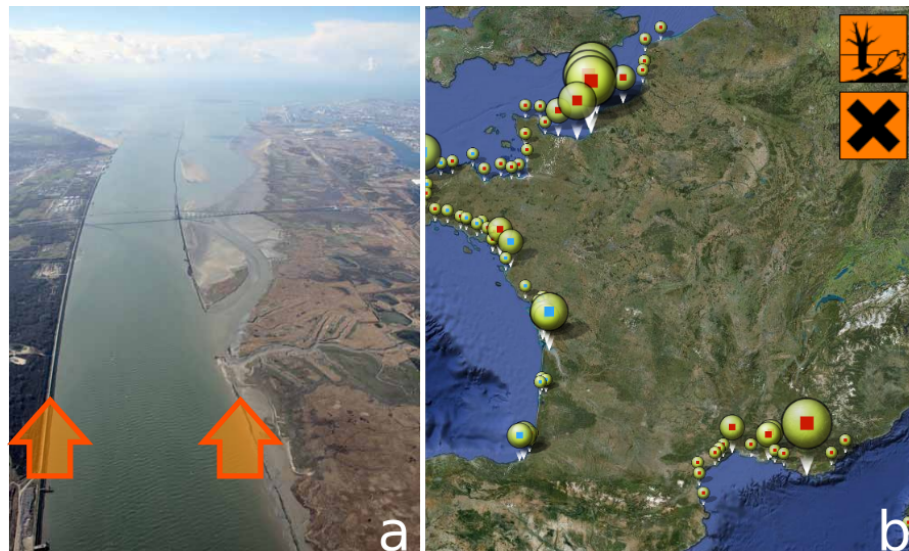


Figure 2 – (a) L'estuaire de la Seine. (b) Niveau de contamination chimique le long des côtes françaises (Ifremer, 2011)

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   tapes
 - Mod  le statistique reliant les densit  s aux covariables
 - Pr  dire les habitats potentiels

2.3. Donn  es

Campagne standardis  e de chalut    perche dans la baie de Vilaine (Fig. 3)

- 1984 – 2010
- En automne
- Juv  niles de l'ann  e (  ge 0)
 - Nb individus / 1000m²

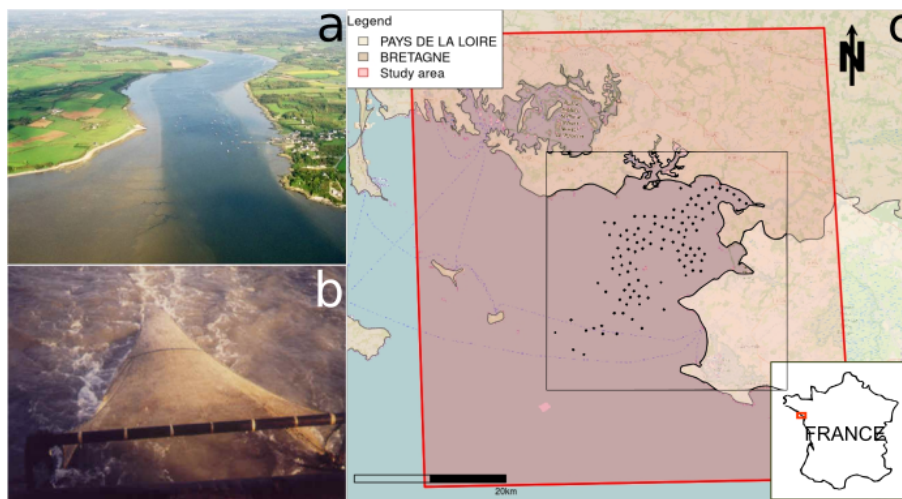


Figure 3 – (a) L'estuaire de la Vilaine. (b) Chalut    perche. (c) Situation des stations d'  chantillonnage.

2.4. Covariables

- Bathym  trie (Fig. 4a)
 - MNT    1000m de r  solution
 - Projection Mercator
- Structure s  dimentaire (Fig. 4b)
 - Fichier shape de polygones
 - Coordonn  es g  ographiques
- Zones biologiques (Fig. 4c)
 - Combinaison bathym  trie, s  diment, habitat
 - Fichier shape de polygones
 - Coordonn  es g  ographiques

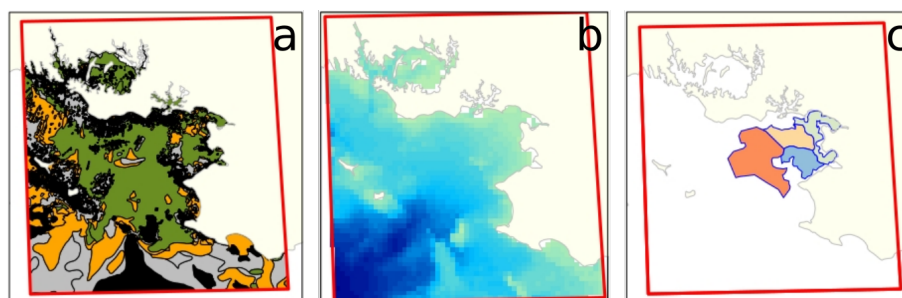


Figure 4 – Covariables en baie de Vilaine. (a) Structure s  dimentaire, (b) Bathym  trie et (c) Zones biologiques.

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

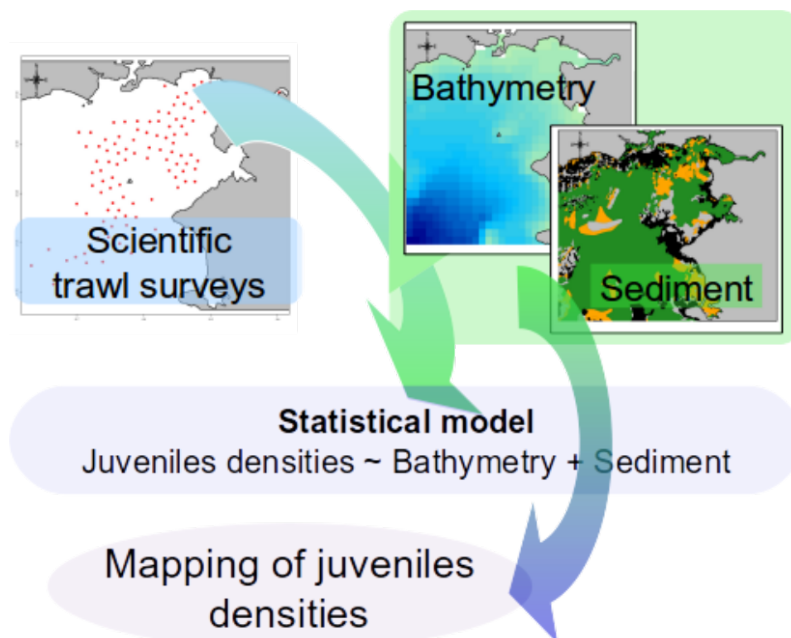


Figure 5 – Procédure pour un modèle de distribution d'espèce

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_Teacher.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 Exploration

L'utilisation d'une transformation log des données montre des distributions proche d'une loi Gaussienne lorsqu'on sépare par covariables (Fig. 6).

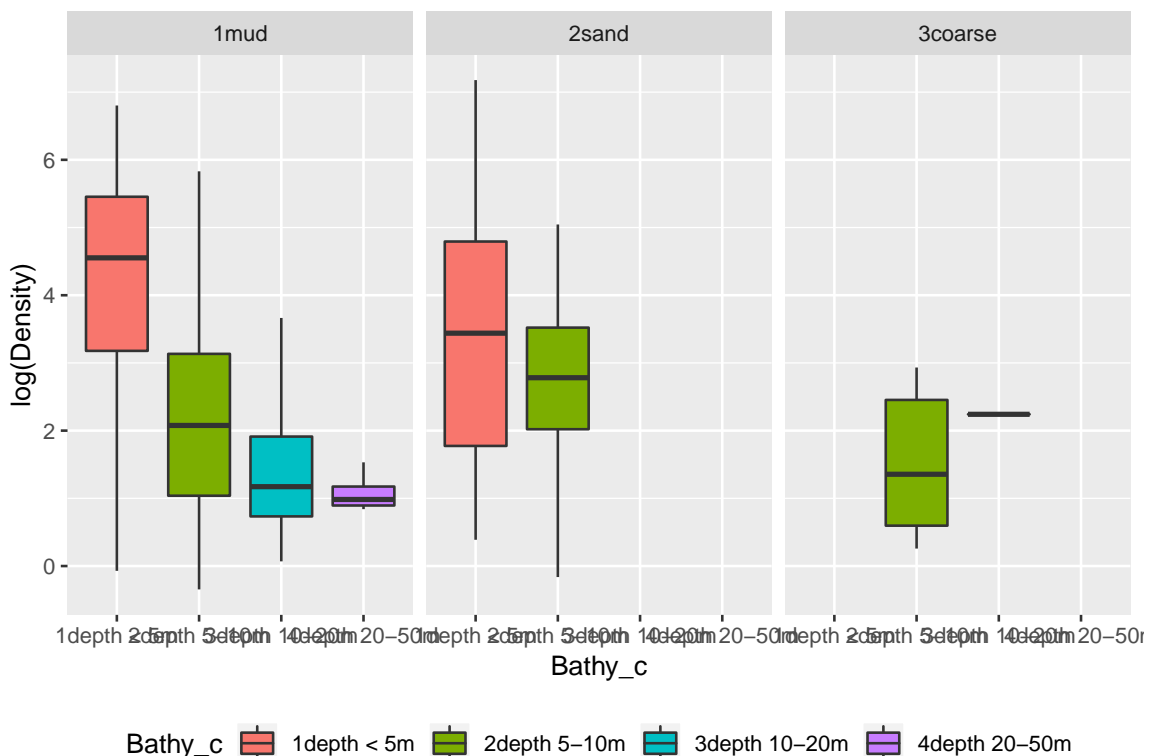


Figure 6 – Observations des densités log-transformées par rapport aux covariables Bathymétrie et Sédiments

4.2.3 Sélection du meilleur modèle

L'exploration des données et le critère d'Akaike conduisent à choisir une distribution log-normale pour les données. En travaillant avec des covariables continues (comme ici la bathymétrie), il est possible que la relation avec les observations ne soit pas linéaire. Dans le cadre des GLM, vous pouvez utiliser des polynômes de différents degrés pour intégrer la non-linéarité de la réponse. L'analyse des résidus sur le modèle sélectionné montre des résultats plutôt satisfaisant (Fig. 7).

4.2.4 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

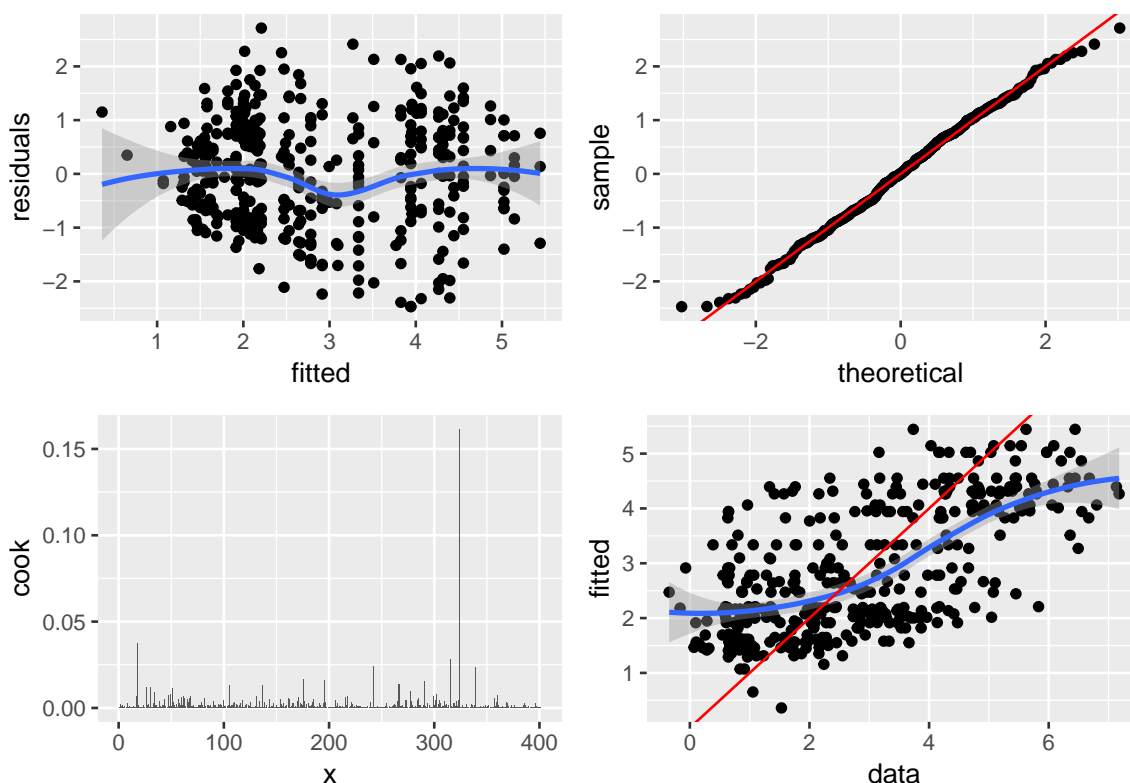


Figure 7 – Figures de diagnostic du meilleur mod  le s  lectionn  

Vous pourriez utiliser le coefficient de cor  lation entre les pr  dictions et les donn  es de validation comme un indice de qualit   d'ajustement pour s  lectionner le meilleur mod  le. Cependant, l'erreur quadratique moyenne ($MSE = \text{Mean Squared Error}$) voire sa racine ($RMSE = \text{Root MSE}$) est l'indice recommand  . Il mesure la distance moyenne d'une observation    sa pr  diction.

La validation crois  e en k -parties est une des meilleurs fa  ons de faire de la validation crois  e. La validation crois  e en $k = 10$ parties est l'une des plus utilis  es. Elle divise le jeu de donn  es en 10 parts   gales et r  p  te la validation crois  e pour chacune des 10 sous-parties utilis  es comme jeu de donn  es de validation (Fig. 8).

Dans notre cas, la validation crois  e est un peu d  licate car nous avons des r  p  titions d'observations sur chaque station   chantillonn  e plusieurs ann  es de suite. Si la variabilit   inter-annuelle est faible, toutes les donn  es d'une m  me station seront   gales et donc les donn  es de validation seront similaires aux donn  es d'ajustement, rendant la validation crois  e peu int  ressante. *Soyez donc prudents avec la validation crois  e lorsqu'il y a suspicion de forte cor  lation de vos donn  es !* Pour passer outre ce probl  me de cor  lation, il faut s  lectionner les donn  es de validation de mani  re judicieuse...

- *Le mod  le s  lectionn   sur la base de l'AIC est-il toujours le meilleur mod  le avec le RMSE ?*

4.3. Sous-mod  le Binomial

4.3.1   tapes

La proc  dure    adopter avec le sous-groupe de donn  es est la m  me qu'avec le jeu de donn  es complet.

- Cr  er les observations de pr  sence-absences    partir du jeu de donn  es
- Explorer ce nouveau jeu de donn  es (Fig. 9)
- Utiliser une distribution binomiale
 - Tester les covariables, les interactions, les fonctions de lien, les crit  res de qualit  
- Choisir le meilleur mod  le

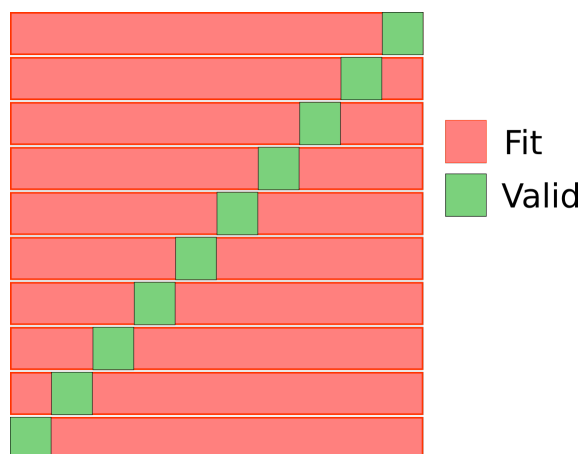


Figure 8 – Illustration de la s  lection de jeux de donn  es de validation pour une validation crois  e en 10 parties

4.3.2 Exploration

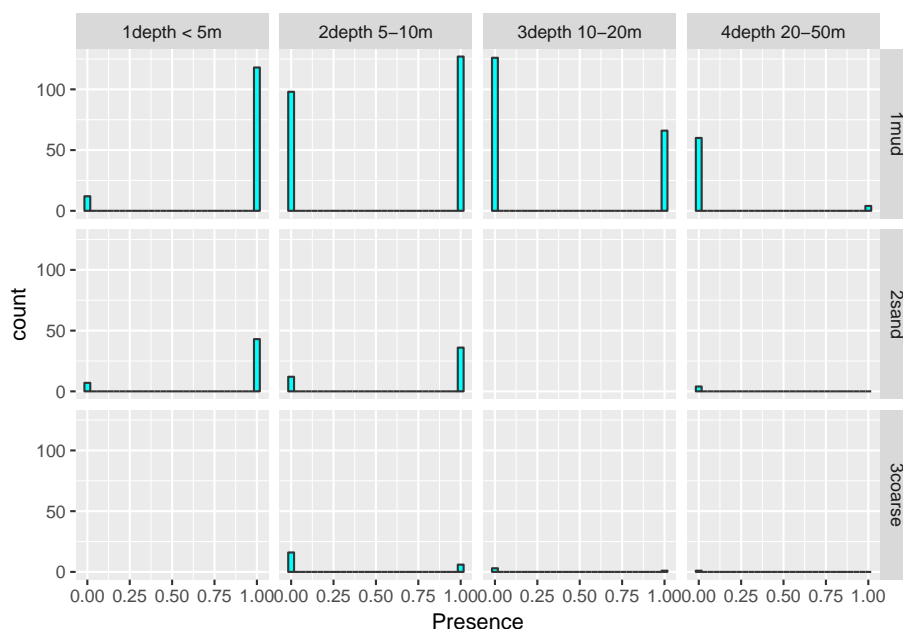


Figure 9 – R  partition des observations en fonction de la bathym  trie et des s  diments

4.3.3 Ajuster un mod  le binomial avec une fonction de lien

Le choix de la distribution pour un mod  le de pr  sence-absence est simple, c  est un mod  le binomial. Cependant, un mod  le est g  n  ralement ajust   sur la base de r  sidus Gaussiens. Pour ajuster un mod  le binomial, les donn  es doivent   tre transform  es de telle sorte qu  on puisse ajuster un mod  le lin  aire Gaussien classique dessus. Pour cela, nous utilisons une fonction de lien. La fonction de lien classique d  un mod  le binomial est la fonction logit, mais ce n  est pas la seule. Vous pouvez tester cloglog, probit ou cauchit.

La fonction logit est la suivante (Fig. 10):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Cette fonction tranforme les valeurs dans l  intervalle [0;1] en valeurs dans $[-\infty; \infty]$, de telle sorte que le mod  le ajust   soit :

$$\text{logit}(p) = \text{Covariate1} + \text{Covariate2} + N(0, \sigma)$$

où p est la probabilité de présence que l'on peut retrouver après ajustement en utilisant la fonction inverse (logit^{-1}).

L'analyse des résidus d'un modèle binomial est aussi à faire, même si on n'a pas vraiment le choix du modèle. Les sorties graphiques sont particulières à analyser (Fig. 11).

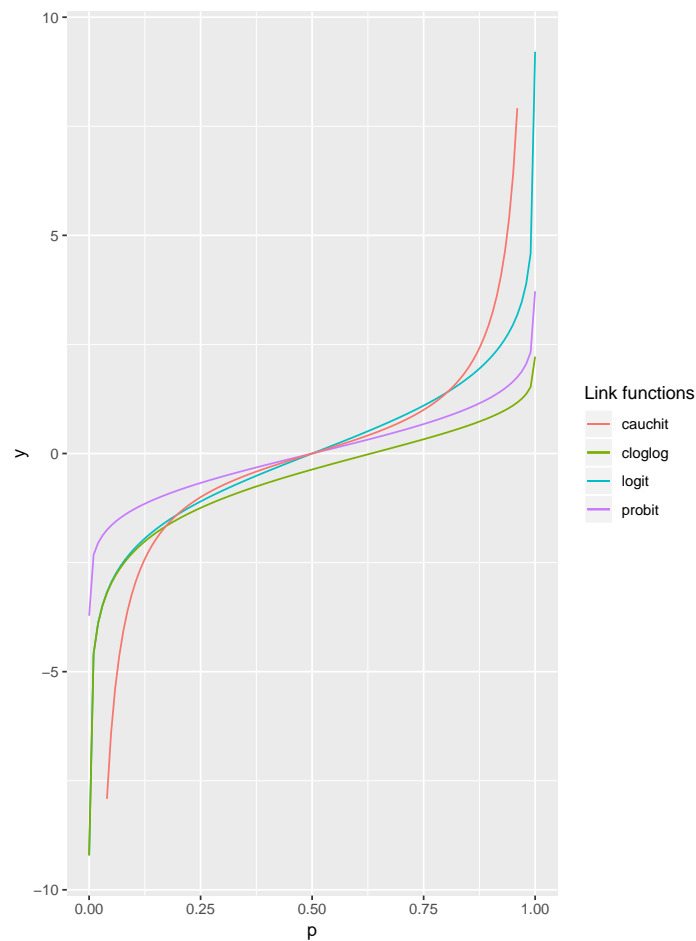


Figure 10 – Différentes fonctions de lien possibles pour un modèle binomial

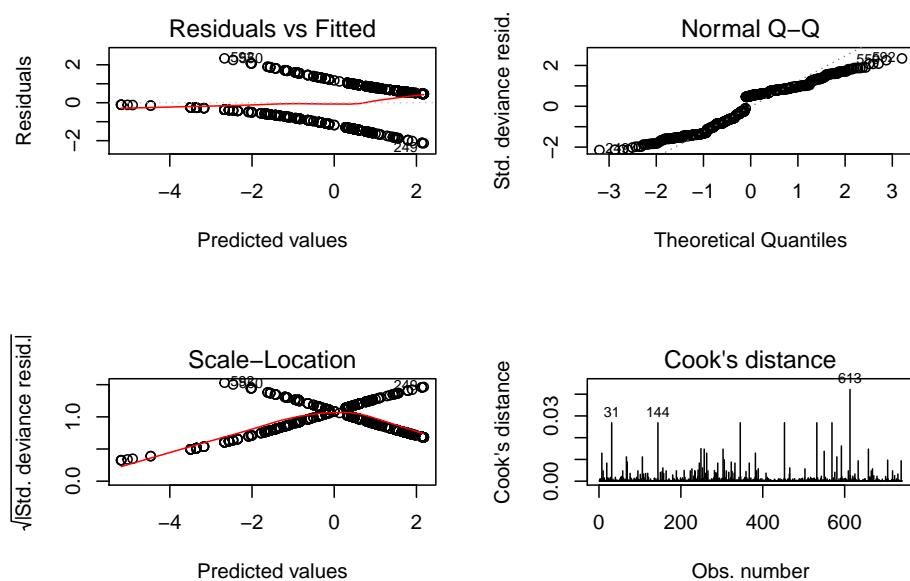


Figure 11 – Analyse des résidus d'un modèle binomial

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.

L'AUC s'appelle ainsi parce qu'elle est calculée à partir d'une courbe "ROC" (Receiving Operating Characteristic) qui compare le taux de vrais positifs (sensitivity) au taux de faux positifs (specificity) pour différentes valeurs de seuil (Fig. 12).

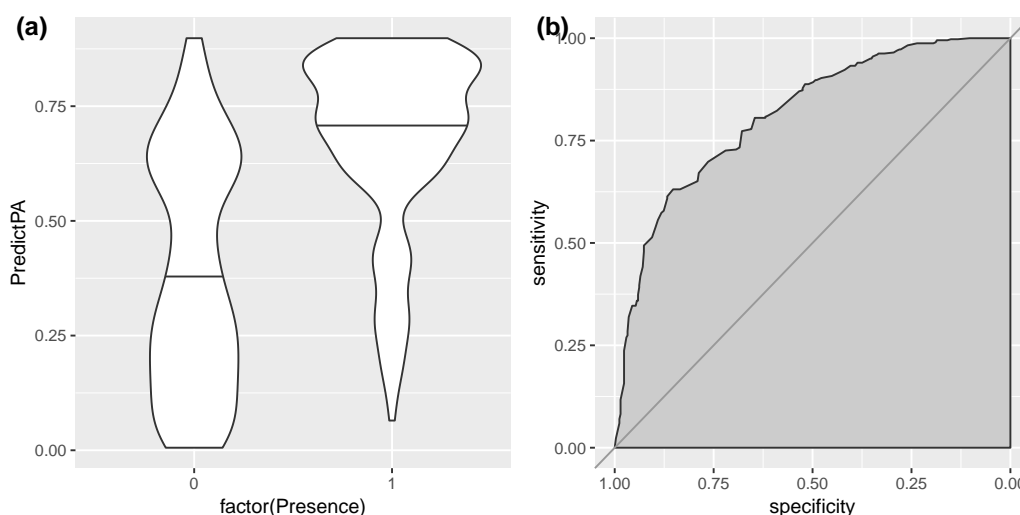


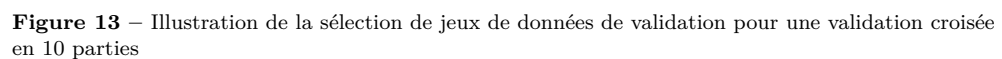
Figure 12 – (a) Prédiction vs Observations. (b) Courbe ROC d'un modèle binomial

4.3.5 Choix du meilleur seuil

La validation croisée en k -parties est une des meilleurs façons de faire de la validation croisée. La validation croisée en $k = 10$ parties est l'une des plus utilisées. Elle divise le jeu de données en 10 parts égales et répète la validation croisée pour chacune des 10 sous-parties utilisées comme jeu de données de validation (Fig. 8).

Dans notre cas, la validation croisée est un peu délicate car nous avons des répétitions d'observations sur chaque station échantillonnée plusieurs années de suite. Si la variabilité inter-annuelle est faible, toutes les données d'une même station seront égales et donc les données de validation seront similaires aux données d'ajustement, rendant la validation croisée peu intéressante. *Soyez donc prudents avec la validation croisée lorsqu'il y a suspicion de forte corrélation de vos données !* Pour passer outre ce problème de corrélation, il faut sélectionner les données de validation de manière judicieuse...

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



11 / 17

de covariables. Cette moyenne ne représente pas la variabilité des observations, ni la proportion d'absences, ni les quelques fortes valeurs de densités observées. Pour prédire parfaitement la variabilité des observations, il faudrait intégrer les covariables environnementales qui l'explique. Le modèle développé ici n'intègre pas la totalité des effets environnementaux, mais il donne une bonne idée des densités moyennes observées en fonction des covariables sélectionnées.

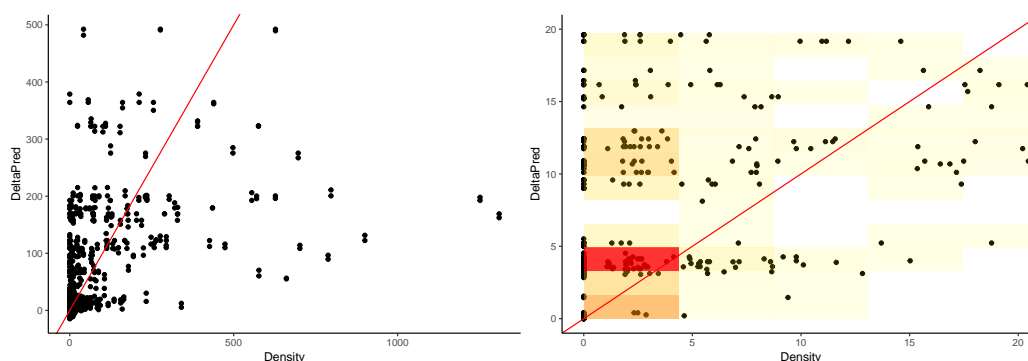


Figure 15 – Prédictions comparées aux observations. La figure de droite est un zoom de celle de gauche.

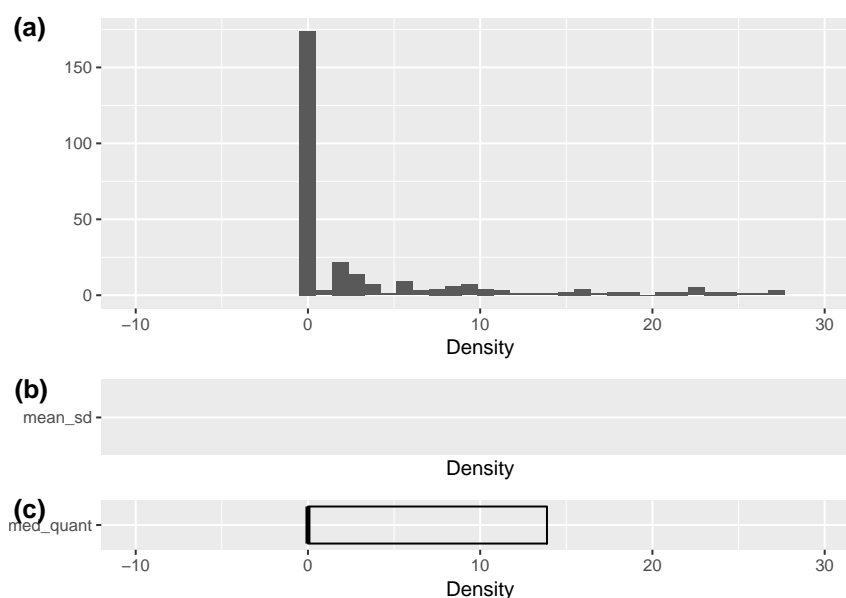


Figure 16 – (a) Forme de distribution issue d'un modèle Delta. Représentation de l'incertitude (b) moyenne et 2*écart-type, (c) médiane et quantiles 10% - 90%.

4.4.3 Prédictions

Le modèle peut être utilisé pour prédire sur un nouveau jeu de données (Fig. 17). Avec ces modèles, il est très simple de calculer des prédictions pour n'importe quelle valeur des covariables, cependant, il est important de ne jamais prédire en dehors de l'étendue des valeurs de covariables observées (Fig. 18). Le nombre d'observation dans les différentes combinaisons de covariables a aussi de l'importance, c'est pourquoi l'exploration des données est nécessaire. De plus, en utilisant une transformation log des données, une petite différence de prédiction peut devenir très élevée dans l'échelle d'origine des données.

4.5. Conclusion

- Séparation des absences et des données positives
 - Construction d'un modèle adapté aux données binomiales

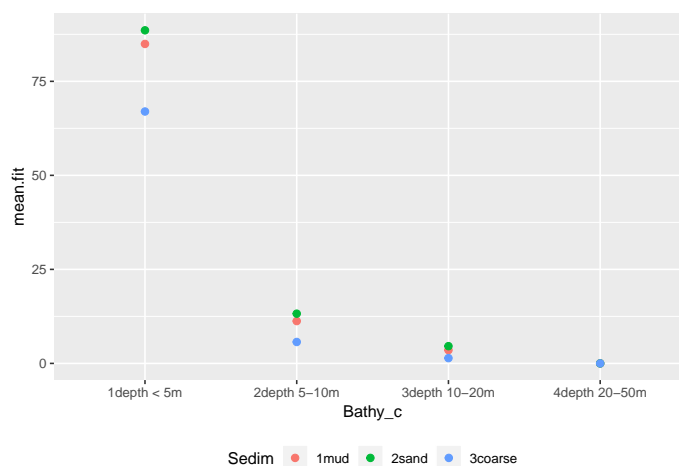


Figure 17 – Prédiction des moyennes pour les différentes combinaisons des covariables

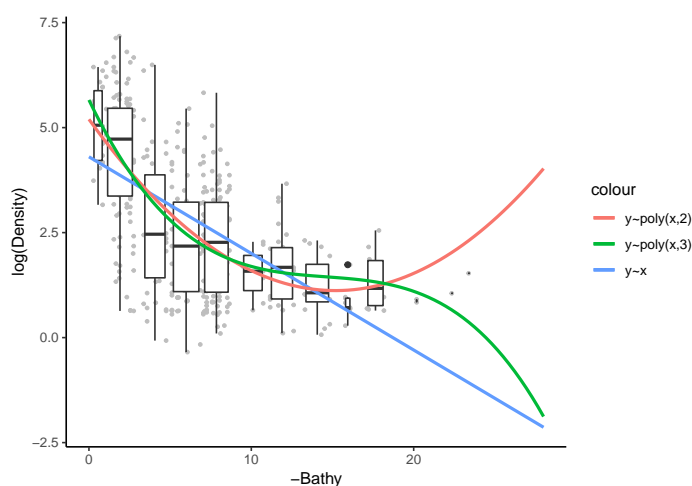


Figure 18 – Illustration de modèles ajustés sur les densités log-transformées et leurs extrapolations

- Construction d'un modèle adapté aux données positives à large distribution
- Qualité d'ajustement sur les sous-modèles meilleure que sur les données brutes
- Couplage des deux sous-modèles
 - Interprétation biologique sensée
 - Question de l'estimation de l'incertitude
- Prédications : Indice de qualité des habitats

5. Modèle d'habitat

5.1. Étapes

La réalisation d'une carte de distribution d'espèce (Fig. 19) 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

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. 20).

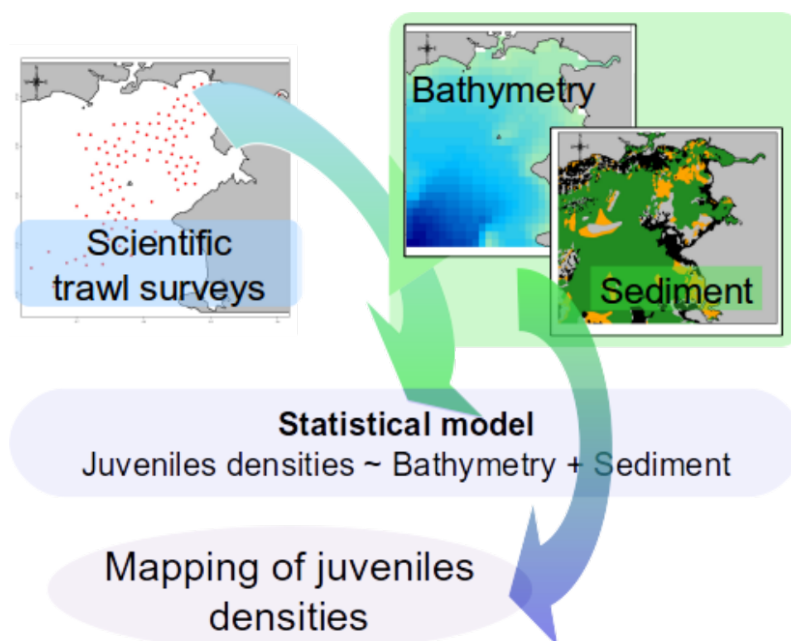


Figure 19 – Procédure pour cartographier une distribution d'espèce

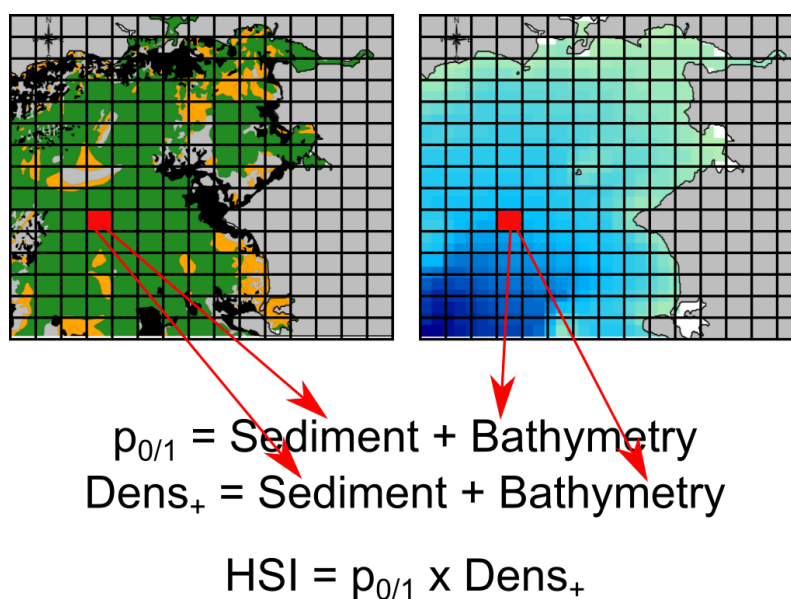


Figure 20 – Prédiction de densité pour chaque cellule d'une carte au format raster

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 niveaux 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. Prédications

La fonction `predict` a une méthode pour pouvoir être utilisée directement sur un objet Raster (Fig. 21).

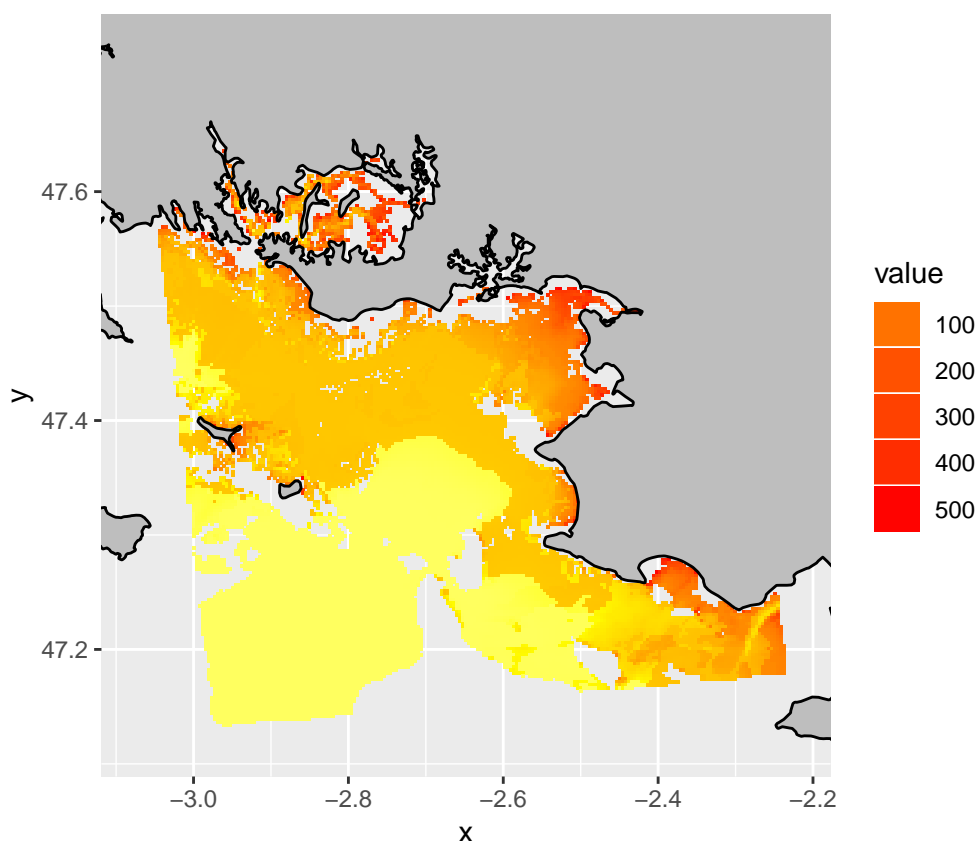


Figure 21 – Prédiction des densités de soles en baie de Vilaine. Échelle de couleur en fonction des quantiles des données originales (10%, 50%, 75%, 95%).

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)

- 16 / 17

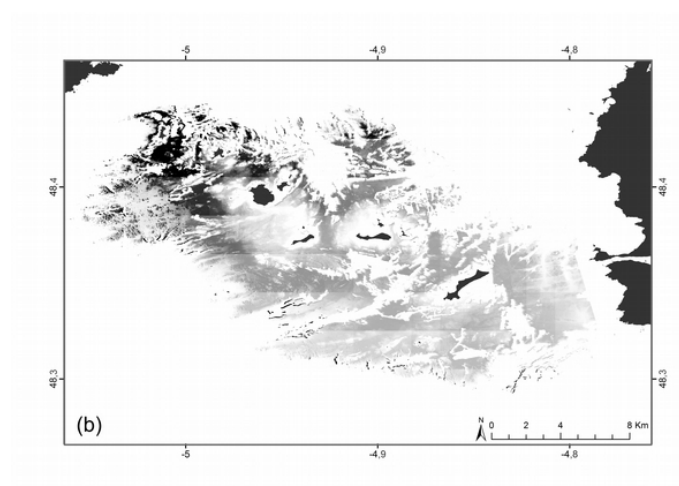


Figure 23 – Estimation spatialisée des biomasses de laminaires dans le parc marin d'Iroise pour la gestion de la ressource. Bajjouk T., Rochette S., Ehrhold A., Laurans M., Le Niliot P. (2015). Multi-approach mapping to help spatial planning and management of the kelp species *L. digitata* and *L. hyporborea*: Case study of the Molène archipelago, Brittany. *Journal of Sea Research*.