



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	1
2.3	Donn��es	2
2.4	Covariables	3
2.5	Ajuster un mod��le de distribution d��esp��ces	3
2.6	Exploration des donn��es	3
3	Pr��paration	4
3.1	Structure des dossiers	4
3.2	D��butons avec R	4
3.3	Sous-mod��le Binomial	5

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

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.

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
 - Perte de surface disponibles (Fig. 2a)
 - Qualit   des habitats alt  r  e (Fig. 2b)
- Impact sur le renouvellement des populations
 - Jeune stades = Gouleau d'  tranglement
 - La taille et la qualit   des nourriceries c  ti  res influent sur la production de juv  niles

Objectifs

D  terminer les facteurs ayant une influence sur la distribution des poissons plats (*Solea solea*) en Baie de Vilaine et cartographier la distribution moyenne des densit  s.

- Cartographier les habitats potentiels n  cessite:
 - Connaissance des habitats de juv  niles
 - Campagnes d'  chantillonnage dans la zone d'  tude
 - Connaissance des covariables environnementales ayant potentiellement de l'influence

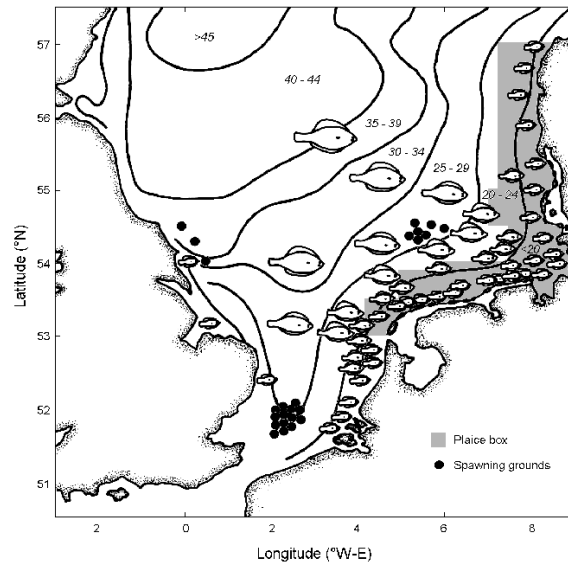


Figure 1 – Plaise box (Rijnsdorp *et al.*)

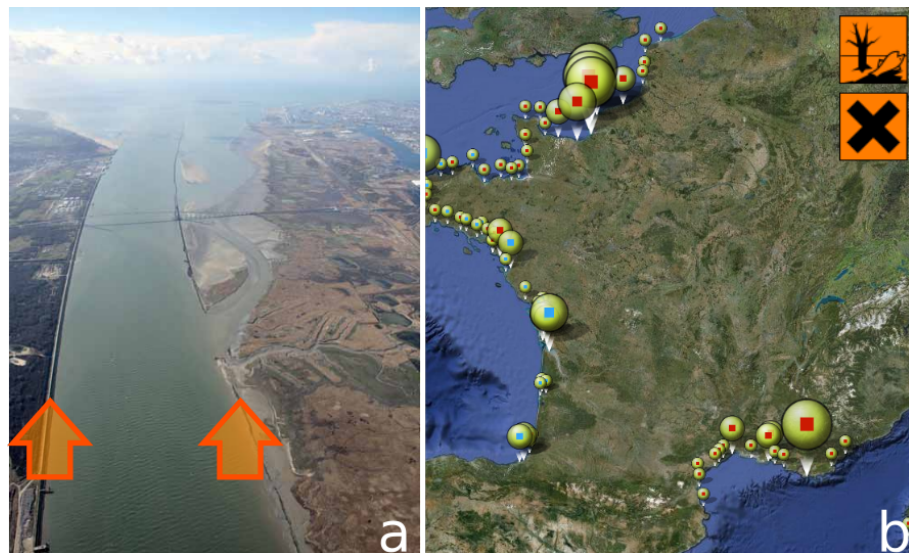


Figure 2 – (a) L'estuaire de la Seine. (b) Niveau de contamination chimique le long des côtes françaises (Ifremer, 2011)

- Cartes exhaustives des covariables environnementales
- Une approche statistique en deux étapes
 - Modèle statistique reliant les densités aux covariables
 - Prédire les habitats potentiels

Données

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

- 1984 – 2010
- En automne
- Juvéniles de l'année (Âge 0)
 - Nb individus / 1000m²

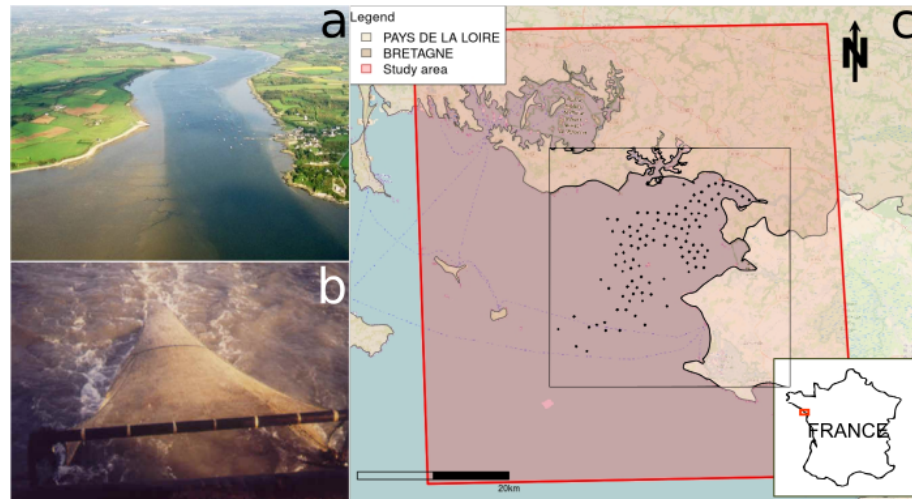


Figure 3 – (a) L'estuaire de la Vilaine. (b) Chalut à perche. (c) Situation des stations d'échantillonnage.

Covariables

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

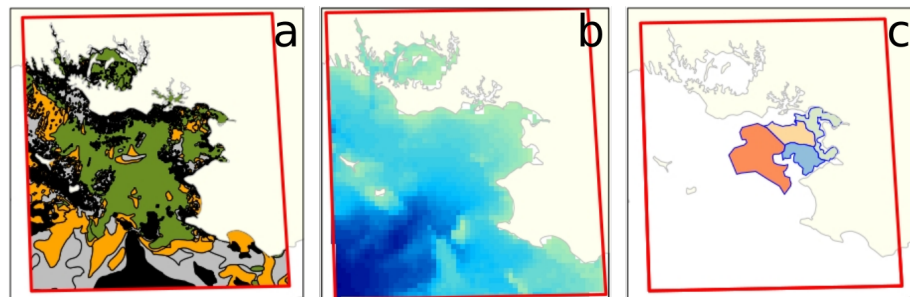


Figure 4 – Covariables en baie de Vilaine. (a) Structure sédimentaire, (b) Bathymétrie et (c) Zones biologiques.

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

Exploration des données

Prenez le temps d'explorer vos données avant toutes analyses

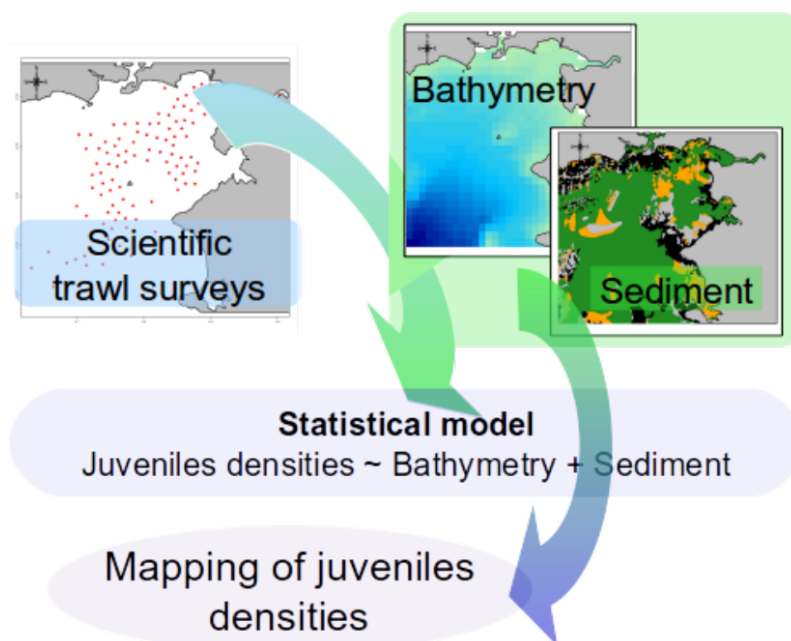


Figure 5 – Procédure pour un modèle de distribution d'espèce

- 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 ?

Préparation

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

Débutons avec R

- Créer un projet Rstudio dans le dossier principal de travail.

- Ouvrez le script R : "Quick_PresAbs_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")
```

Sous-modèle Binomial

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 (Fig. 6)
- Utiliser une distribution binomiale
 - Tester les covariables, les interactions, les fonctions de lien, les critères de qualité
- Choisir le meilleur modèle

Exploration

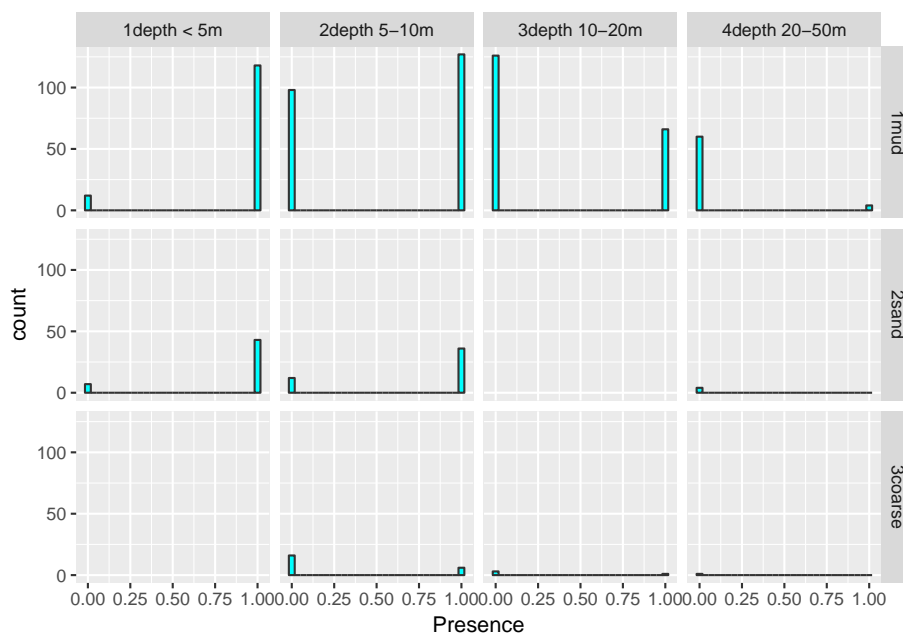


Figure 6 – Répartition des observations en fonction de la bathymétrie et des sédiments

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. Cependant, 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 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. 7):

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

Cette fonction transforme les valeurs dans l'intervalle $[0;1]$ en valeurs dans $[-\text{Inf};\text{Inf}]$, de telle sorte que le mod  le ajust   soit :

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

o   p est la probabilit   de pr  sence que l'on peut retrouver apr  s ajustement en utilisant la fonction inverse (logit^{-1}).

L'analyse des r  sidus d'un mod  le binomial est aussi    faire, m  me si on n'a pas vraiment le choix du mod  le. Les sorties graphiques sont particuli  res    analyser (Fig. 8).

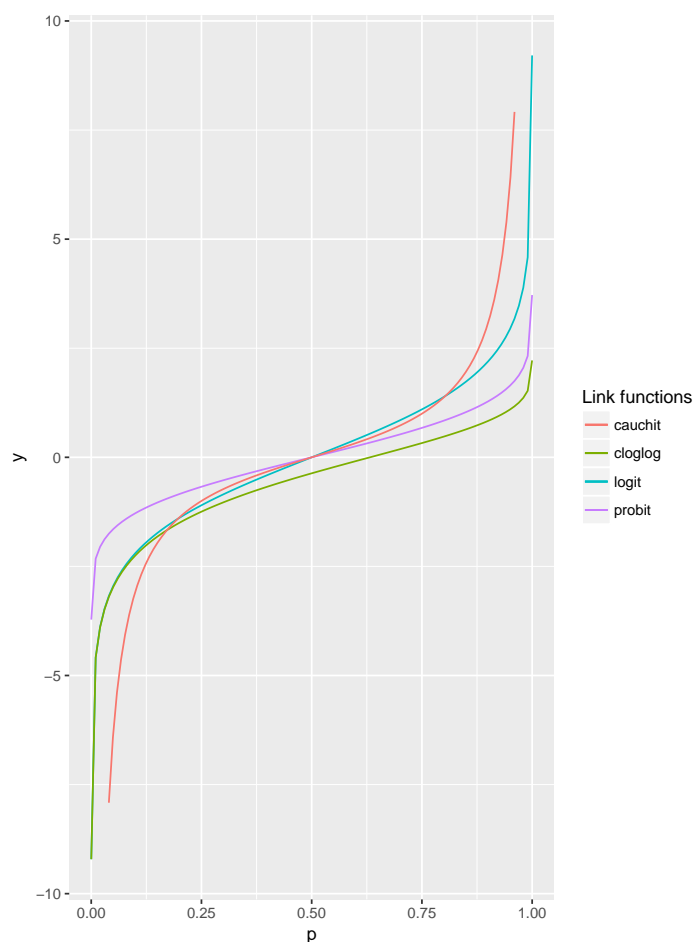


Figure 7 – Diff  rentes fonctions de lien possibles pour un mod  le binomial

Qualit   d'ajustement d'un mod  le binomial

Une mesure couramment utilis  e pour la qualit   d'ajustement d'un mod  le binomial est "l'aire sous la courbe" (AUC : Area Under the Curve). Un objectif des mod  les binomiaux   tant de pr  dire un succ  s ou un   chec, et non pas seulement une probabilit   de succ  s, on peut vouloir d  finir un seuil (intuitivement 0.5 par exemple) qui transforme la probabilit   de pr  sence en pr  sence ou absence. L'AUC est en quelque sorte une probabilit   de classer correctement les pr  sences et absences. Une d  finition plus compl  te serait :

La probabilit   moyenne pour qu'une observation=1 et une observation=0 choisies de mani  re al  atoire dans le jeu de donn  es montrent une probabilit   de pr  sence pr  dite sup  rieure pour l'observation=1 par rapport    celle de l'observation=0

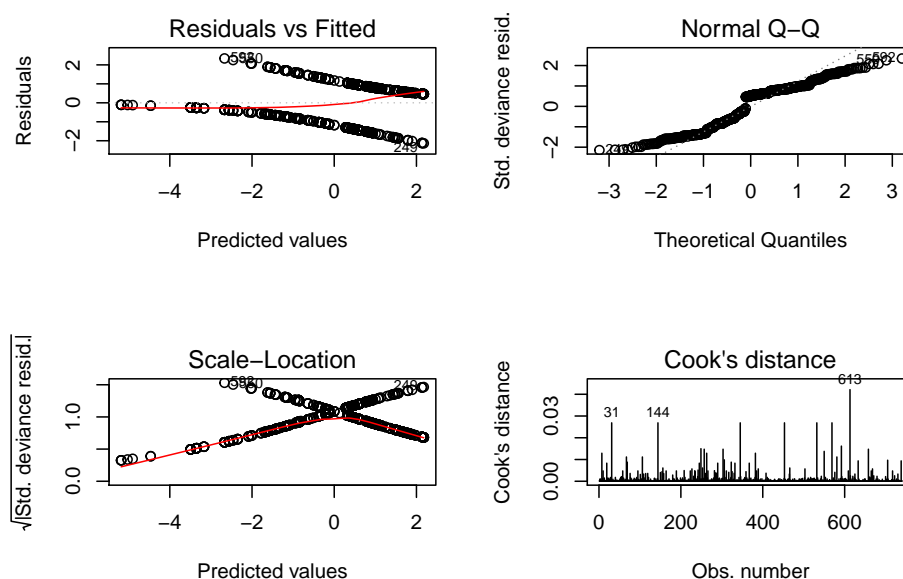


Figure 8 – Analyse des r  sidus d'un mod  le binomial

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

L'AUC s'appelle ainsi parce qu'elle est calcul  e    partir d'une courbe "ROC" (Receiving Operating Characteristic) qui compare le taux de vrais positifs (sensitivity) au taux de faux positifs (specificity) pour diff  rentes valeurs de seuil (Fig. 9).

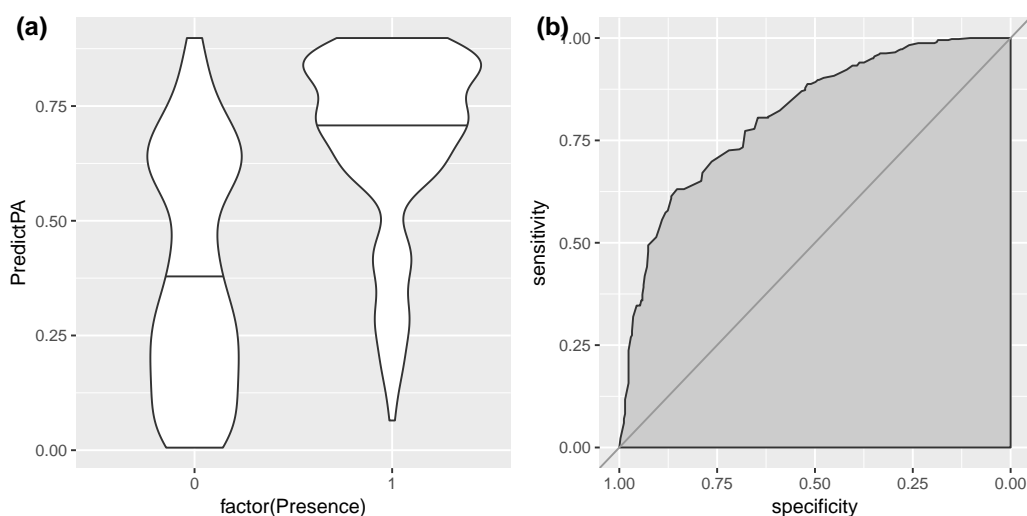


Figure 9 – (a) Pr  diction vs Observations. (b) Courbe ROC d'un mod  le binomial

Choix du meilleur seuil

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. 10).

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 corr  lation de vos donn  es !* Pour passer outre ce probl  me de corr  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 l'AUC sur les donn  es de validation ?*

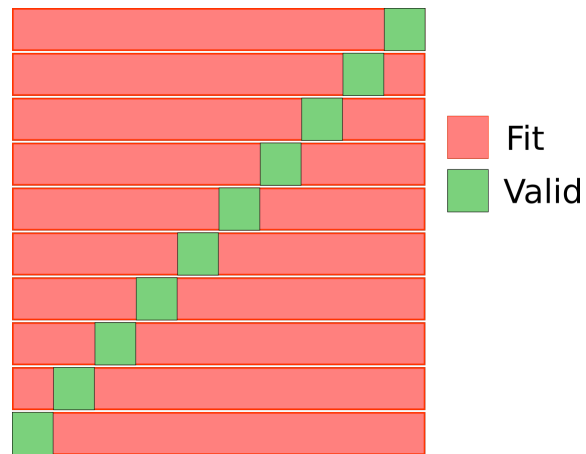


Figure 10 – Illustration de la s  lection de jeux de donn  es de validation pour une validation crois  e en 10 parties