



Formation à R

Modélisation avec les GLM

SÉBASTIEN ROCHETTE, THINKR

ThinkR

1 / 6

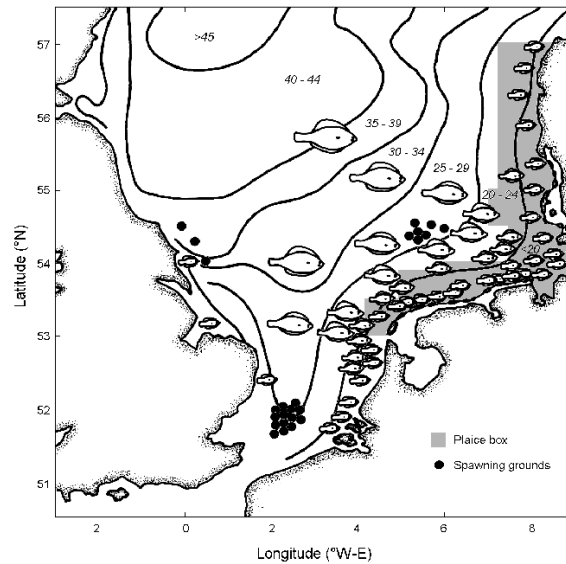


Figure 1 – Plaise box (Rijnsdorp *et al.*)

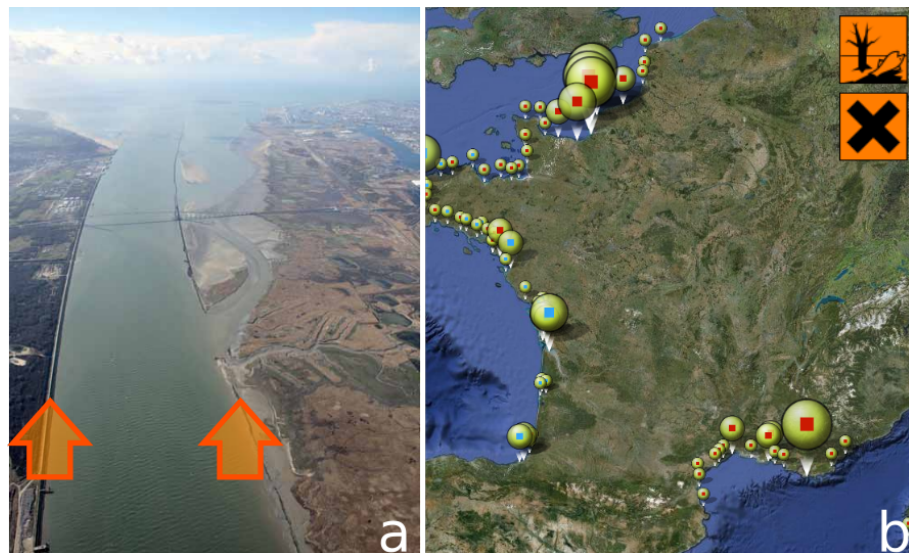


Figure 2 – (a) L'estuaire de la Seine. (b) Niveau de contamination chimique le long des côtes françaises (Ifremer, 2011)

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

2.3. 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

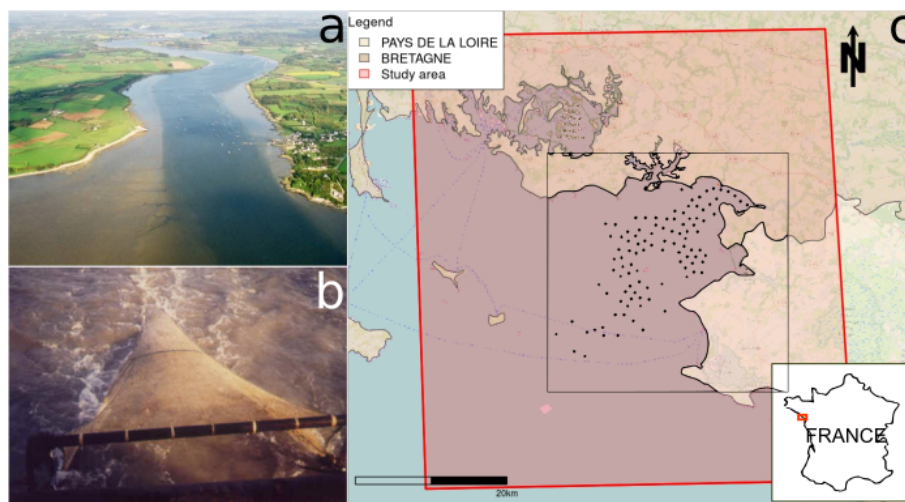


Figure 3 – (a) L'estuaire de la Vilaine. (b) Chalut à perche. (c) Situation des stations d'échantillonnage.

2.4. 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

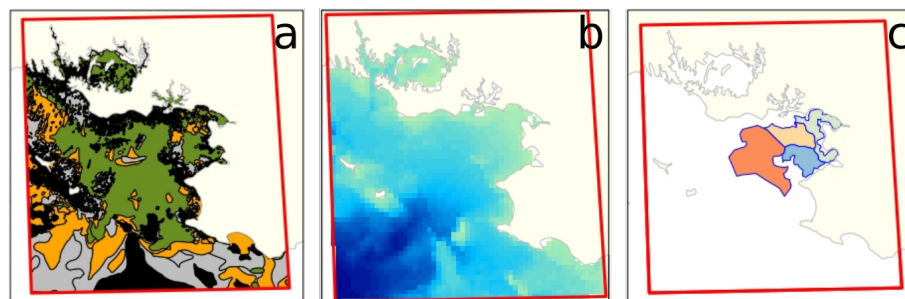


Figure 4 – Covariables en baie de Vilaine. (a) Structure sédimentaire, (b) Bathymétrie et (c) Zones biologiques.

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

2.6. Exploration des données

Prenez le temps d'explorer vos données avant toutes analyses

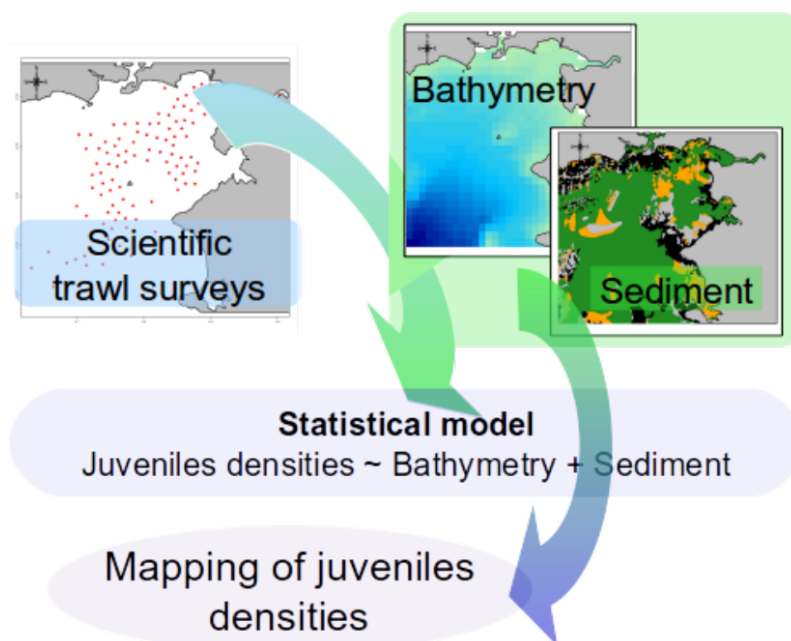


Figure 5 – Procédure pour un modèle de distribution d'espèce

- 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 ?

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. Débuts avec R

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

- Ouvrez le script R : "Quick_PresAbs_Student.R"
- Lister les diff  rents sous-dossier de travail au d  but 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")
```

3.3. Sous-mod  le Binomial

3.3.1   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
- Utiliser une distribution binomiale
 - Tester les covariables, les interactions, les fonctions de lien, les crit  res de qualit  
- Choisir le meilleur mod  le

3.3.2 Exploration

3.3.3 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.

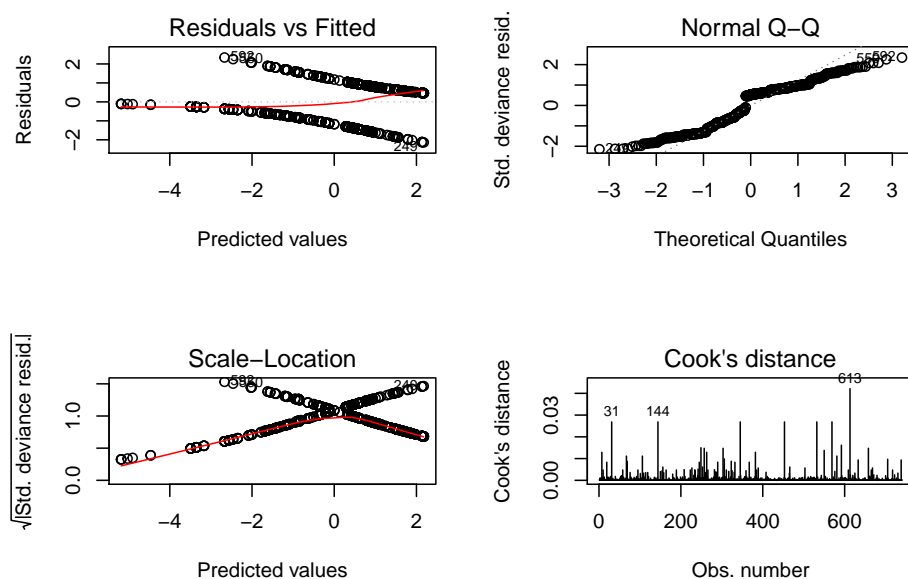


Figure 6 – (ref:RFigOutputsPACap)

3.3.4 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

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

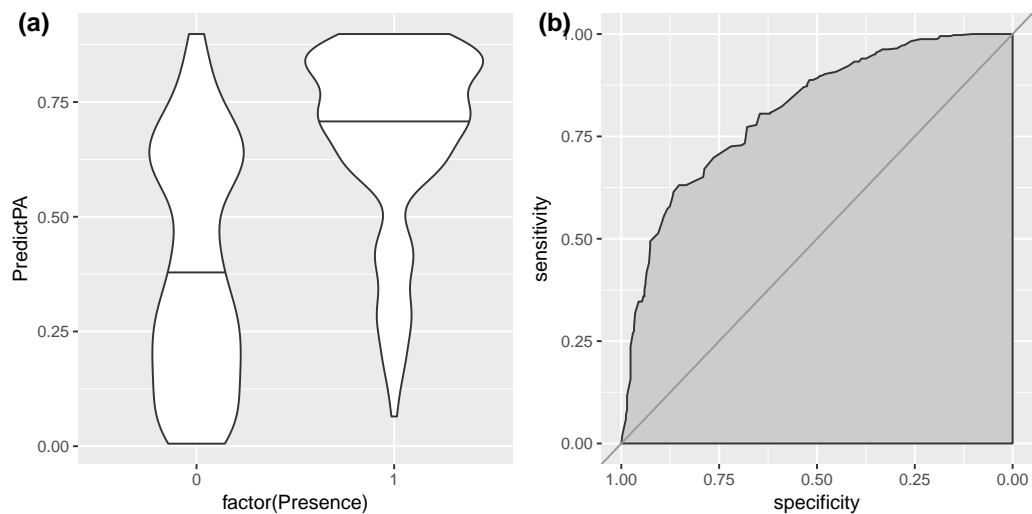


Figure 7 – (ref:RFigAUCPACap)

3.3.5 Choix du meilleur seuil

3.3.6 Validation du mod  le

De m  me que pour le mod  le sur les donn  es positives, vous pouvez utiliser la validation crois  e en k parties pour s  lectionner le meilleur mod  le en terme de pr  diction. Pour un mod  le binomial, vous pouvez utiliser l'AUC sur le jeu de donn  es de validation comme un indice de qualit   d'ajustement.