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 - Teacher
```

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)
  - 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

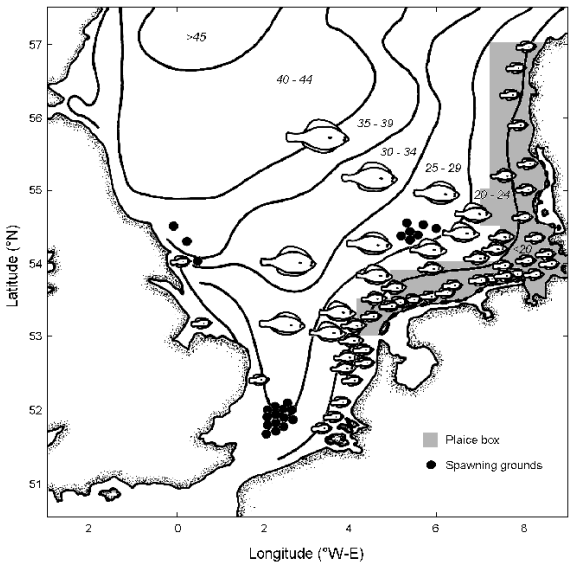

Figure 1 – Plaice box (Rijnsdorp *et al.*)

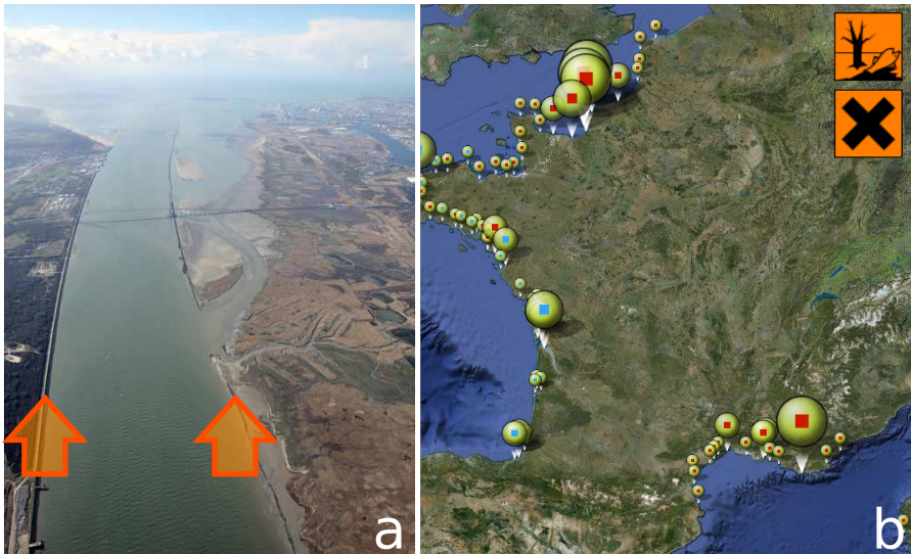

Figure 2 – (a) L'estuaire de la Seine. (b) Niveau de contamination chimique le long des côtes françaises (Ifremer, 2011)

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

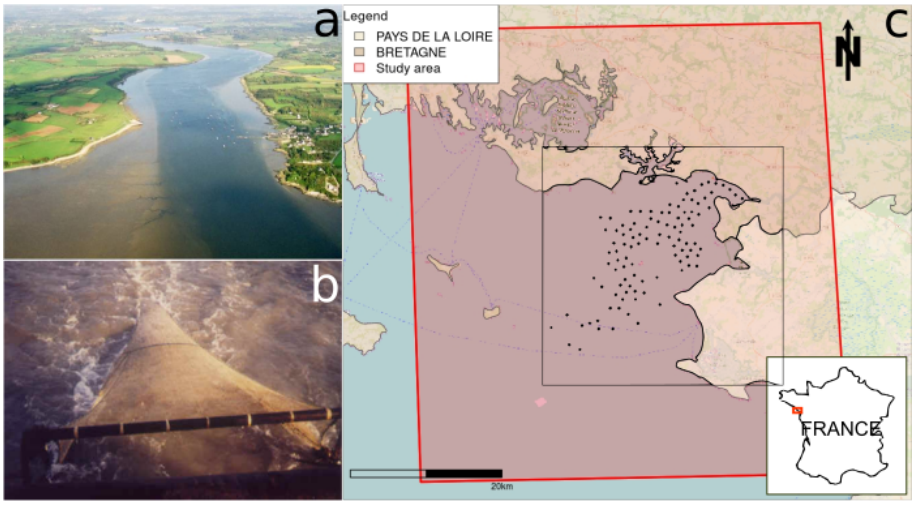

Figure 3 – (a) L'estuaire de la Vilaine. (b) Chalut à perche. (c) Situation des stations d'échantillonnage.

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

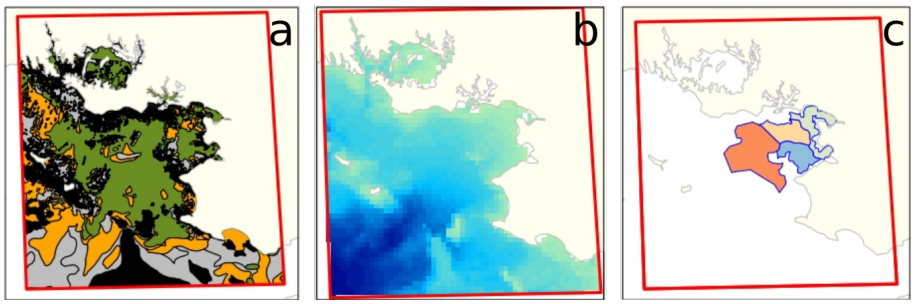

Figure 4 – Covariables en baie de Vilaine. (a) Structure sédimentaire, (b) Bathymétrie et (c) Zones biologiques.

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

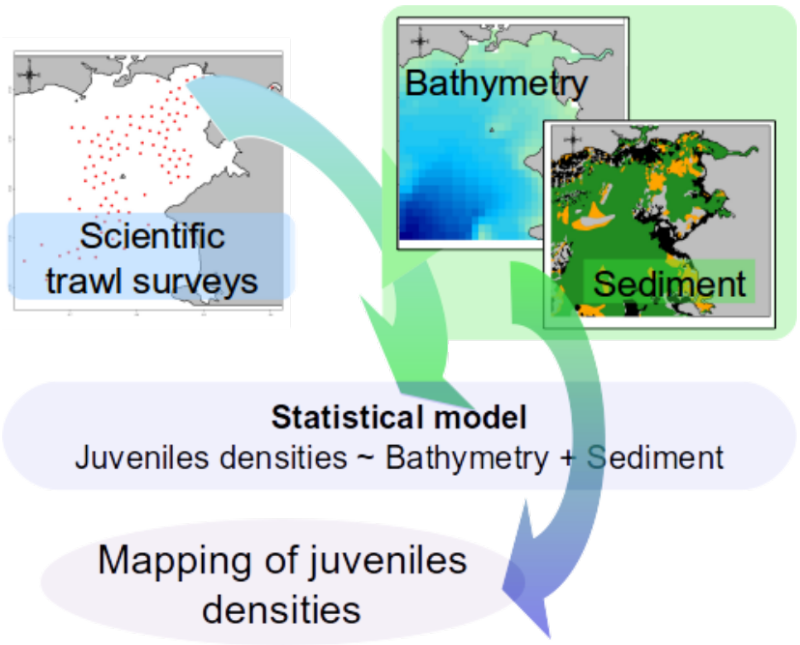

Figure 5 – Procédure pour un modèle de distribution d'espèce

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_AllDataModel_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. Exploration des données

4.1. Étapes

Souvenez-vous : Définissez ce que vous cherchez, à quelles questions vous souhaiteriez répondre !

- Explorer la répartition du plan d'échantillonnage en fonction des covariables environementales
- Explorer les données d'observation au regard des covariables environnementales pour détecter de potentielles corrélations
- Explorer les interactions entre les effets des covariables sur les observations
- Explorer les lois de distribution possibles (gaussian, log-normal, ...) des observations en fonctions des combinaisons de covariables

Les scripts qui sont fournis ne sont que des exemples, ils ne son
solutions ! Faîtes vos propres tests !

4.2. Liste des différentes étapes

- Lire le jeu de données spatialisé (Fig. 6)
- Ajouter une nouvelle covariable : la bathymétrie divisée en classes
  - " $<$ 5 m","5-10 m","10-20 m","20-50 m"
- Explorer la répartition des observations en fonctions des covariables
  - Centrer l'analyse sur l'année, la bathymétrie et le sédiment
  - *Que remarquez-vous ?*
- Explorer les covariables ayant potentiellement des effets sur les densités
  - *Quelles covariables pourraient avoir une influence ?*(Fig. 7)

Les modèles statistiques que nous allons utiliser peuvent se résumer de cette façon :

$$Density = Covar1 + Covar2 + Noise$$

Comme vous le savez, on chercher toujours à savoir si les données sont gaussienne pour pouvoir procéder à l'analyse statistique. Si elles ne sont pas gaussienne, nous devons définir le type de distribution pour pouvoir utiliser une transformation de données.

- Explorer la distribution des données
  - *Quelle est la distribution la plus intéressante ?*

La figure 8 montre différents exemples de distributions.

Exemple de l'effet de deux facteurs (Fig. 9)

- *Qu'en pensez-vous ?*
- Explorer les interactions potentielles entre les covariables

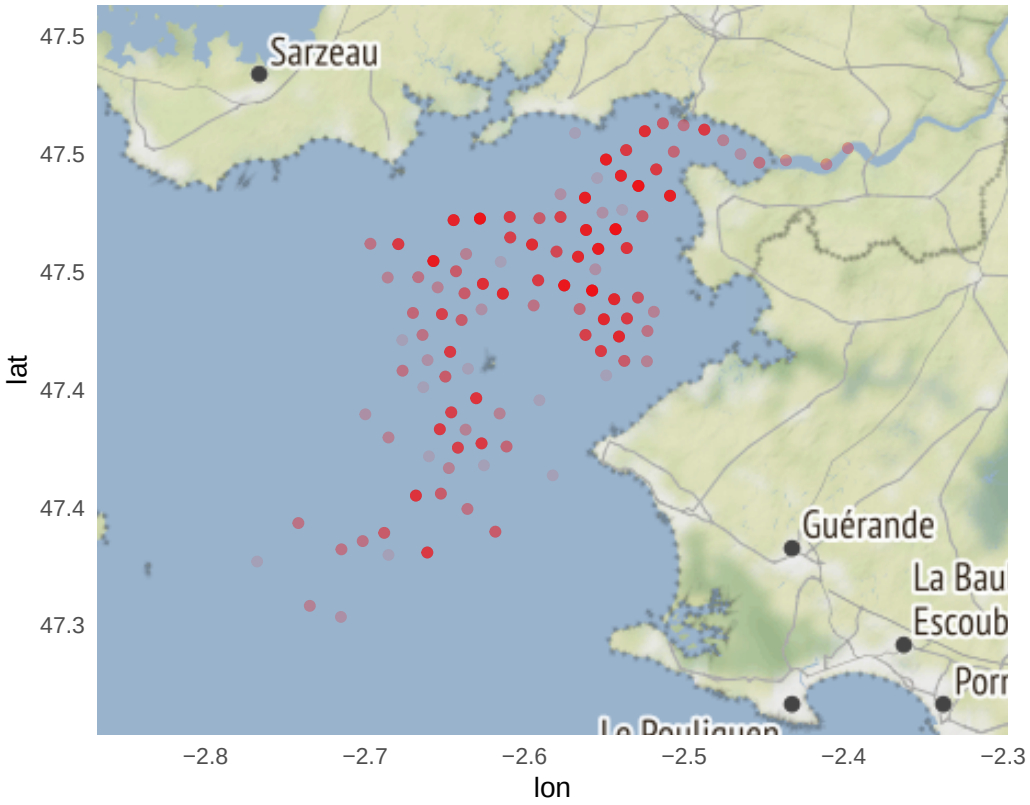

Figure 6 – Répartition des stations d'échantillonnage

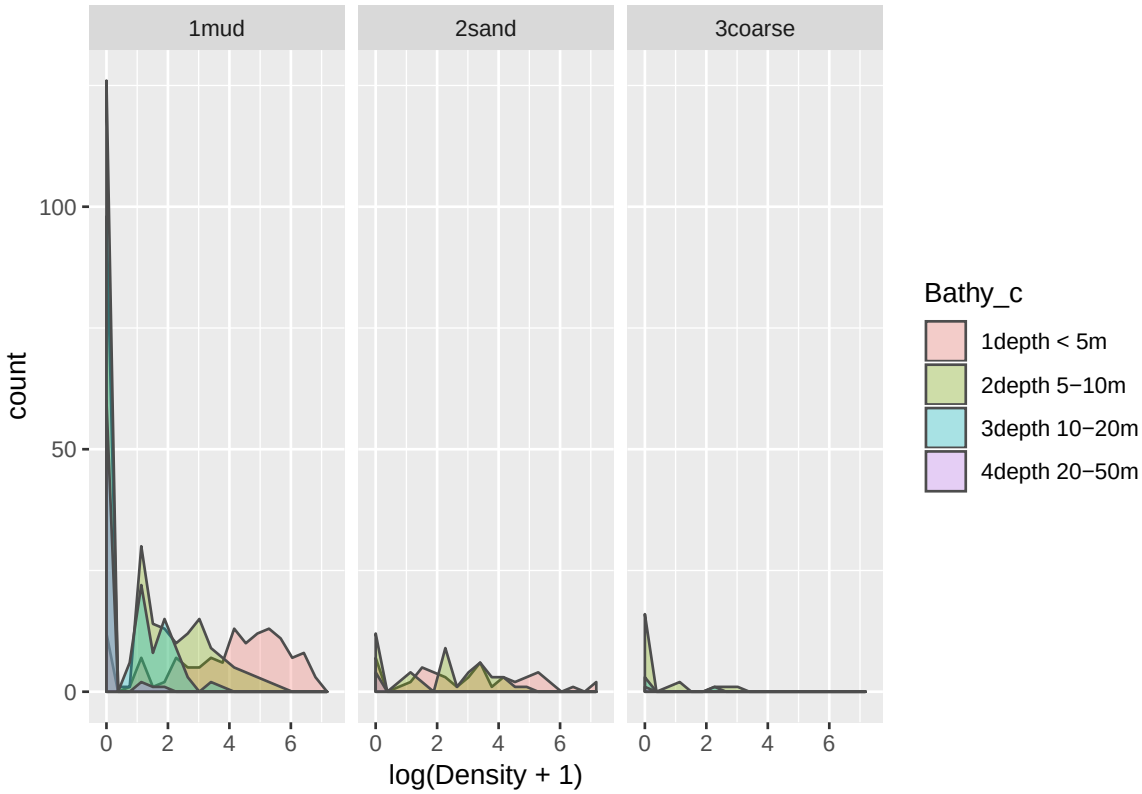

Figure 7 – Densités (log-transformées) en fonction de la bathymétrie et des sédiments

- *Qu'en pensez-vous ?*
- Comme aide à l'interprétation, utiliser l'exemple théorique de la figure 10

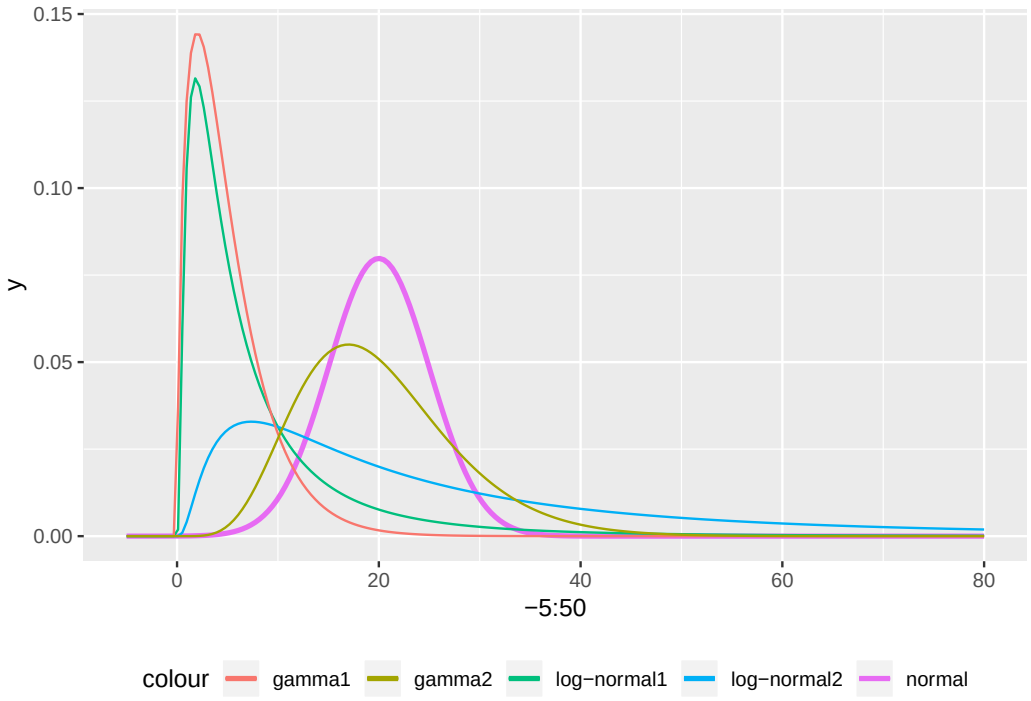

Figure 8 – Différents exemples de distributions

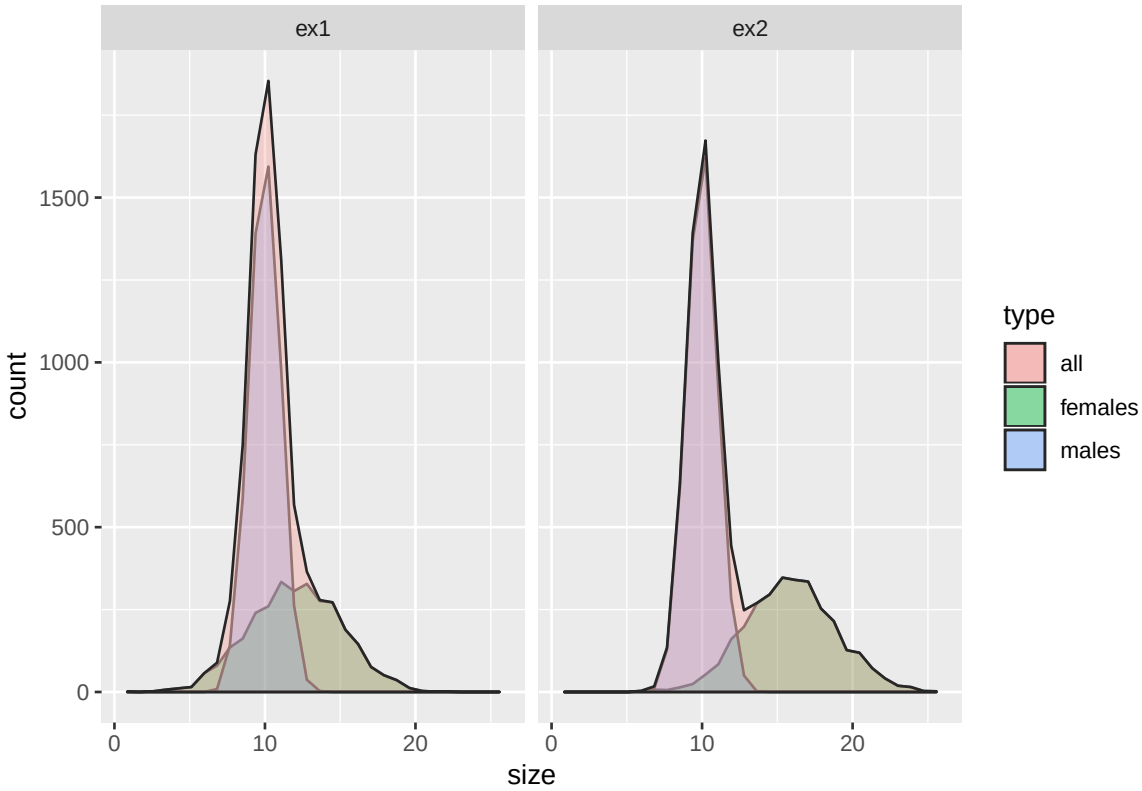

Figure 9 – Differents effets de deux facteurs

5. Modélisation

5.1. Étapes

Souvenez-vous: Définissez ce que vous cherchez, à quelles questions vous souhaiteriez répondre !

- Tester les différentes formes de modèles au regard des combinaisons de covariables et des

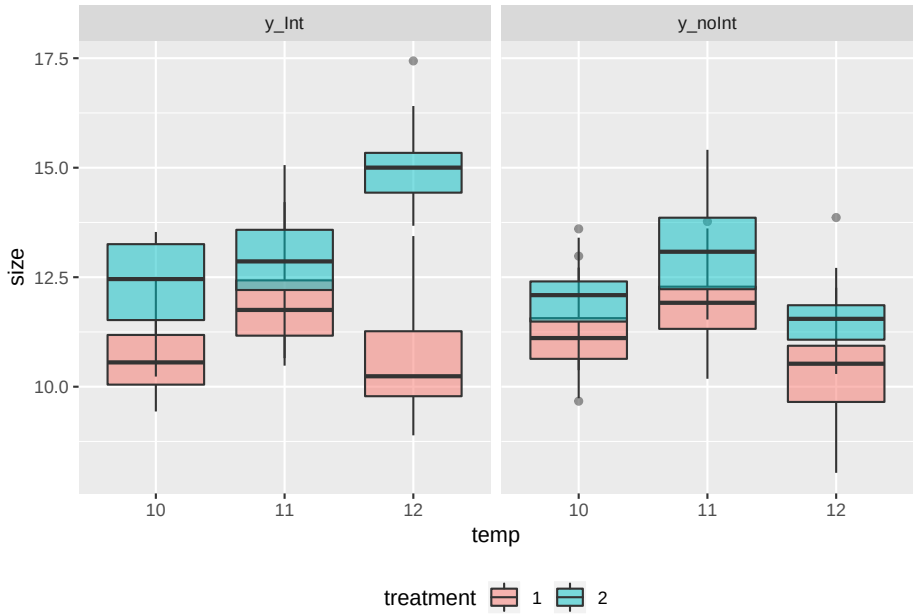

Figure 10 – Interaction entre traitements et température

formes de distributions des résidus

- Comparer les modèles à l'aide des outils statistiques à disposition (AIC, anova, ...), de la validation croisée mais aussi de la connaissance du jeu de données et des questions ciblées
- Analyser les résidus des modèles. Analyser leur distribution et s'assurer que les hypothèses de construction sont vérifiées.
  C'est seulement lorsque les hypothèses sur la distribution des résidus sont vérifiées, que les covariables et les interactions sélectionnées peuvent commencer à être interprétées...
  Les scripts qui sont fournis ne sont que des exemples, ils ne sont en aucun cas les meilleures solutions ! Faîtes vos propres tests !

5.2. Interpréter les sorties de modèles

Lorsque vous ajustez un modèle linéaire (lm ou glm), vous pouvez utiliser différents tests statistiques et visuels qui répondent à différentes questions. Votre question principale pourrait être :

- "Est-ce que mes covariables ont un effet sur mes observations ?". En réalité, ce n'est pas exactement la question à laquelle va répondre votre modèle. Ce serait plutôt "Est-ce que les covariables que j'ai utilisées expliquent une part de la variabilité de mes observations ?"

Pour que vous puissiez interpréter les différentes sorties de modèles, dans ce document, nous allons regarder le modèle suivant :

$$lm(Density\ Bathy + Sedim, data = dataset)$$

Ce modèle n'est pas forcément le meilleur modèle à choisir !

5.2.1 Summary(lm)

Cette fonction montre un tableau de tests de significativité (Table 1). Ce sont des tests de Student. Ils testent si la valeur estimée pour un effet est ou non significativement différente de zéro. Ainsi, si une covariable a un effet non significativement différent de l'effet nul, il est probablement inutile de la conserver dans le modèle.

- Ici, ce qui est appelé "Intercept" est l'effet de base. Dans une équation $y = a.x+b$, l'"intercept" serait b. Ici, c'est un peu différent car il y a des covariables au format facteur ("Sedim"). Dans cet exemple, l'"intercept" montre une "p-value" proche de zéro, ce qui signifie que son effet (`estimate ~ 100`) est significativement différent de zéro.

Table 1 – Exemple d'une sortie de 'summary(lm)'

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	99.944	8.127	12.298	0.000
Bathy	5.921	0.639	9.260	0.000
Sedim2sand	-0.157	12.938	-0.012	0.990
Sedim3coarse	-41.819	23.241	-1.799	0.072

- La covariable "Bathy" est continue. Dans une équation de type $y = a.x + b$ ce serait a. Dans cet exemple, son effet `estimate ~ 6`, est significativement différent de zéro (`p-value ~ 0`).
- La covariable "Sedim" est un facteur à trois niveaux. Dans le "summary", vous ne pouvez en voir que deux ("2sand", "3coarse"). En réalité, les effets des niveaux de facteur (*estimate*) sont comparés au premier niveau ("1mud"), pour lequel l'effet est égal à zéro. Dans l'équation $Y = a * Bathy + b[Sedim] + c$, l'estimation de la moyenne Y pour `Sedim = "1mud"` est $Y = a*Bathy+0+c$. Dans les résultats du tableau, `b["2sand"]` est donc non significativement différent de `b["1mud"] = 0`, et, `b["3coarse"]` avec une `p-value = 0.07`, n'est pas non plus significativement différent de `b["1mud"] = 0`.

5.2.2 Analyse de résidus

Une hypothèse de construction d'un modèle linéaire est l'homoscédasticité, aussi appelée homogénéité de la variance. Cela signifie que la variance de la réponse `y` est la même quelque soit la valeur du prédicteur `x`. Dans un modèle Gaussien classique ajustant `x` à `y` ainsi: $y = a.x + b + \epsilon$, la variable ϵ représente les résidus du modèle. Ils sont supposés être centrés sur zéro et avec une variance Gaussienne, leur distribution suivant ainsi la loi Gaussienne $\epsilon \sim N(0, \sigma)$
Lorsqu'on simule un tel modèle, par exemple $y = 2.x + 5 + \epsilon$, on observe la figure 11, avec une homogénéité de la distribution des observations autour de l'ajustement, et une distribution Gaussienne des résidus comme le montrent l'histogramme et le "qqplot".

```
# Example of homogeneous residuals
n <- 1000
epsilon <- rnorm(n, 0, 5)
x <- runif(n, 0, 10)
y <- 2*x + 5 + epsilon

par(mfrow = c(1,3))
plot(x, y, pch = 20)
abline(5, 2, col = "red", lwd = 2)
hist(epsilon, breaks = 20, col = "grey")
qqnorm(epsilon); qqline(epsilon, col = "red")
```

À partir de cet exemple, vous pouvez définir les diagnostics graphiques nécessaires pour vérifier vos hypothèses de construction de modèle. Lorsqu'on utilise le même modèle que précédemment (Fig.12) :

- `Residuals vs Fitted` - Dans cet exemple, on peut voir que la variablité des résidus augmente avec les valeurs prédites, ce qui va à l'encontre de l'homogénéité de la variance.
- `Scale-Location` est en accord avec la figure précédente car on voit une augmentation de la deviance des résidus quand les prédictions augmentent.
- `Normal Q-Q` - Cette figure appelée "qqplot" montre la divergence entre les quantiles théoriques d'une loi Gaussienne et les quantiles réels de la distribution des résidus du modèle. La divergence est importante pour les valeurs élevées, ce qui montre une queue de distribution plus longue qu'une loi Normale.
- `Hist. of residuals` - L'histogramme des résidus est en accord avec le qqplot car on voit clairement un distribution qui n'est pas une Gaussienne centrée sur zéro. Cette distribution a une longue queue de distribution avec beaucoup plus de valeurs positives que de valeurs négatives.
  (ref:RFigResidualsDiagCap) Figures de diagnostic d'un modèle linéaire permettant de vérifier les hypothèses de construction.

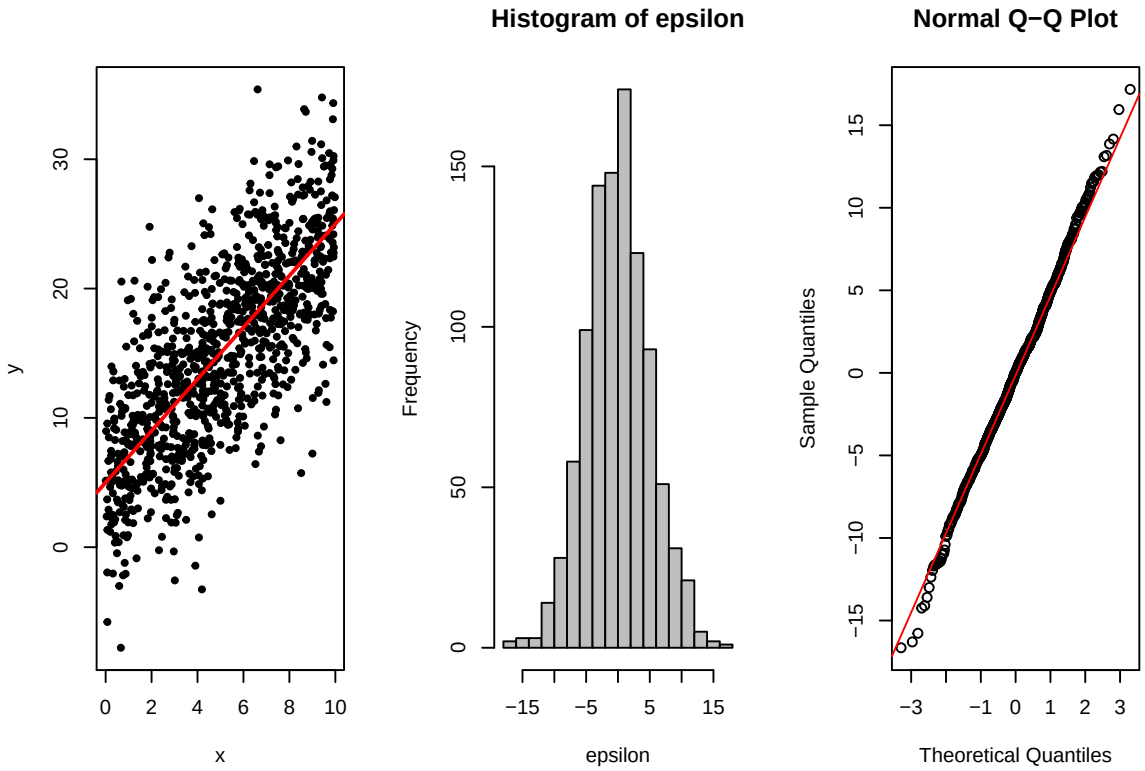

Figure 11 – Simulation d'une relation linéaire entre x et y avec un résidu Gaussien.

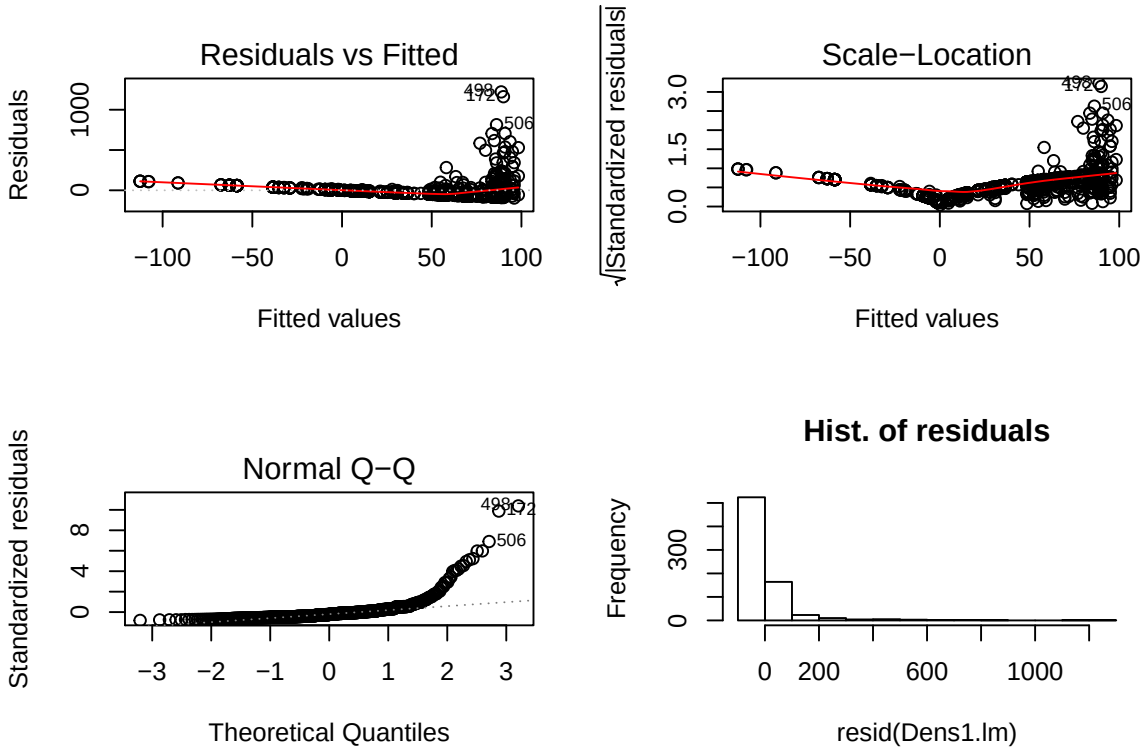

Figure 12 – (ref:RFigResidualsDiagCap)

5.2.3 Analyse de variance

La question à laquelle répond une `anova` (avec un test du Chi-2) est : Est-ce que la covariable ajoutée augmente significativitement la vraisemblance du modèle (ou a réduit la déviance résiduelle), comparé au modèle précédent, sans cette covariable ?

Table 2 – Exemple d'une 'sortie' de anova(lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Bathy	1	1251584	1251584	89.66	0.000
Sedim	2	45447	22724	1.63	0.197
Residuals	736	10274168	13959	NA	NA

Table 3 – Comparaison de la déviance expliquée par différents modèles avec un nombre croissant de paramètres

model	Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)	"||"	df	AIC
lm1	8	3192	NA	NA	NA	||	3	92.0
lm2	6	1272	2	1920	0.003	||	5	86.8
lm3	4	645	2	626	0.144	||	7	84.1

- `NULL` est le test pour un modèle sans covariable, c'est le modèle qui estime la moyenne : $Density \sim constant$.
- `Bathy` est le test pour un modèle uniquement avec la `Bathy` : $Density \sim Bathy$. La `p-value` est proche de zéro, indiquant que le gain de déviance expliquée en ajoutant la `Bathy` est significativement différent de zéro.
- `Sedim` est le test pour un modèle avec `Bathy` et `Sedim`, dans cet ordre : $Density \sim Bathy + Sedim$. La `p-value ~ 0.2` indique qu'il y a un risque de 20% que la déviance expliquée en ajoutant la covariable `Sedim` au modèle contenant déjà la `Bathy` soit nulle.

5.2.4 Critères d'Akaike (AIC) et Bayesian (BIC)

L'AIC et le BIC sont des critères de qualité d'ajustement pénalisés par le nombre de paramètres estimés. La description (traduite) de ces fonctions dans R est :

> Function générique calculant "Le Critère d'Information" d'Akaike pour un ou plusieurs modèles ajustés pour lesquels une "log-vraisemblance" peut être obtenue, en utilisant la formule $IC = -2 * log - likelihood + k * npar$, où `npar` represente le nombre de paramètres estimés, et `k = 2` pour l'AIC classique, ou `k = log(n)` (`n` étant le nombre d'observations) pour le BIC ou SBC (Schwarz's Bayesian criterion).

En effet, plus vous ajoutez de paramètres dans un modèle, plu svous avez de chances que le modèle s'ajuste parfaitement aux données (Table 3, Fig. 13). L'AIC diminue avec la déviance résiduelle et augmente avec le nombre de paramètres ajustés. Plus l'AIC est bas, plus parsimonieux est le modèle.

5.3. Trouver le meilleur modèle

La fonction `lm` n'est utilisée que pour des modèles avec une distribution Gaussienne des résidus. Pour tester d'autres types de distributions, il faut utiliser `glm`, avec un paramètre pour la famille de distribution (`family`). Vous pouvez utiliser des distributions qui autorisent une plus grande queue de distribution que la loi Normale. Parmis les familles disponibles, vous pouvez tester `poisson, quasipoisson, Gamma, Log-gaussian`.

- *Quel est le meilleur modèle au regard des différents critères évoqués ?*

5.3.1 Exploration des sorties du meilleur modèle

- *Que pouvez-vous dire sur le diagnostic complet de votre modèle ?* (Fig. 14)

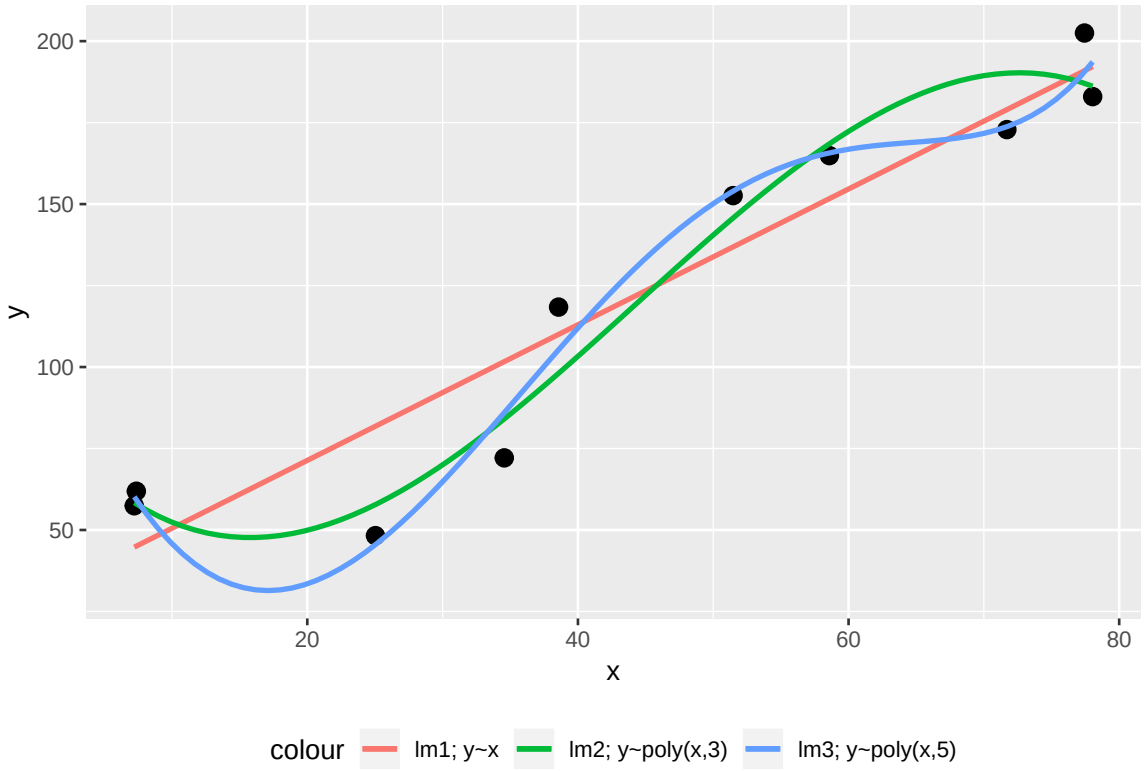

Figure 13 – Différents modèles ajustés sur les mêmes données mais avec un nombre de paramètres ajustés croissant.

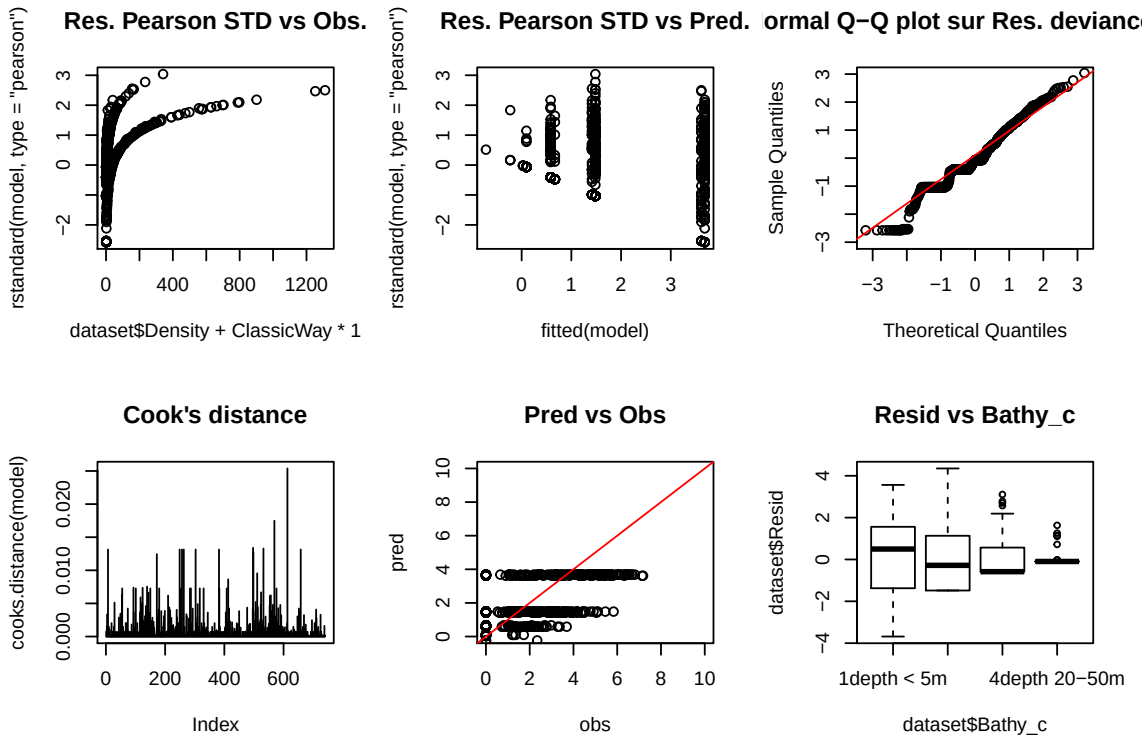

Figure 14 – Diagnostic du meilleur GLM sélectionné

5.4. Prédictions du modèle

Lorsque vous êtes satisfaits du modèle sélectionné, vous pouvez faire des prédictions

- Utiliser le fichier csv fourni pour voir l'effet des covariables sélectionnées (Fig. 15)

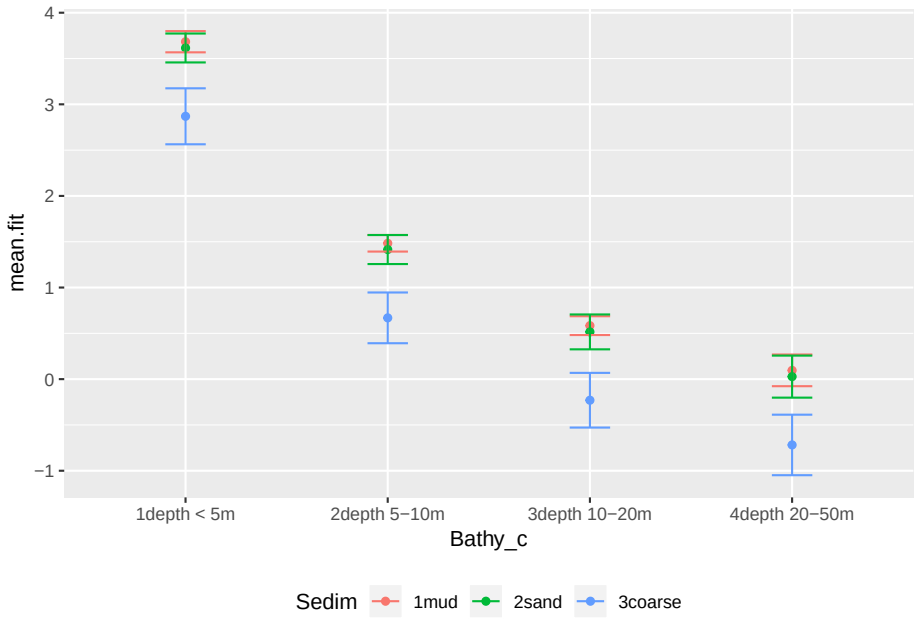

Figure 15 – Prédictions du meilleur GLM sélectionné

5.5. Conclusions sur la modélisation

Le meilleur modèle sélectionné semble donner des résultats intéressants mais on ne doit pas s'en satisfaire car l'analyse des résidus n'est pas du tout satisfaisante.

- Pas de distribution Gaussienne
  - De nombreuses absences
  - Une distribution large avec de fortes valeurs de densités
- Aucune famille de distribution ne satisfait les hypothèses des GLM
  - Résidus Gaussiens et homogénéité de la variance
  - Les familles de loi exponentielles ne sont pas adaptées

Il est nécessaire d'utiliser un modèle qui tient compte de données avec beaucoup d'absences ("zero-inflated data") !