



Université Grenoble Alpes

Master 2 Mathématiques et Applications, parcours Statistique et Science des
Données (SSD)

RAPPORT DE PROJET TUTORÉ

A propos des narvals

Yanis BEN BELGACEM, Vadim BERTRAND, Angélique SAILLET

Encadrés par
Adeline LECLERCQ SAMSON
Frédéric AUDRA

Sommaire

1 Effet des perturbations humaines sur l'alimentation des narvals	2
1.1 Modélisation de l'effet de l'exposition sur le taux d'émission	3
1.1.1 Processus de Poisson	3
1.1.2 Modèle linéaire généralisé	3
1.1.2.1 Effet de la profondeur	4
1.1.2.2 Effet de l'exposition	4
1.1.3 Non-indépendance des observations	4
1.1.3.1 Utilisation de modèles mixtes	4
1.1.3.2 Caractère autorégressif du processus	4
1.1.4 Effet de médiation de la profondeur	5
1.1.5 Intervalles de confiance	6
1.1.5.1 Construction par approche Monte-Carlo	6
1.1.5.2 Estimation de bandes de prédiction via la méthode Delta	8
1.2 Résultats	9
1.2.1 Note sur le temps d'ajustement des modèles	9
1.2.2 Recherche de la mémoire optimale	9
1.2.3 Régression double bi-exponentielle sur les coefficients autorégressifs	10
1.2.4 Impact de l'exposition aux perturbations sur le taux d'émission de buzz	11
1.2.5 Intervalles de confiance	13
1.2.5.1 Coefficients d'exposition	13
1.2.5.2 Pourcentage du taux normal d'émission de buzz	23
Références	24

Partie 1

Effet des perturbations humaines sur l'alimentation des narvals

Notre travail s'inscrit dans une collaboration du Laboratoire Jean Kuntzmann (LJK) de Grenoble avec l'Université de Copenhague et l'Institut groenlandais des ressources naturelles.

Les narvals sont des baleines vivant toute l'année au Groenland. Le réchauffement climatique favorise le recul des glaces sur le territoire groenlandais et ses côtes. Cela ouvre la porte au développement d'activités humaines au Groenland, et notamment des activités minières. Les biologistes de l'Institut groenlandais des ressources naturelles se questionnent sur les effets potentiels engendrés par la présence humaine sur les comportements des narvals.

Afin d'anticiper ces possibles modifications de comportements une étude a été conduite pendant plusieurs mois en 2018 [1]. Dans ce cadre, 6 narvals ont été équipées de capteurs permettant d'enregistrer leur profondeur de plongée, leur localisation et les sons qu'elles émettent. Les baleines ont été laissées libres de toutes perturbations pendant plusieurs jours avant d'y être exposées. Les perturbations ont pris la forme de coups de fusil tirés dans l'eau depuis un bateau afin d'imiter les ondes émises par des activités minières.

Lorsqu'elles se nourrissent, les narvals émettent des sons spécifiques appelés "buzz". À partir des sons collectés il est donc possible de déterminer quand ces baleines sont en train de manger. La distance séparant les baleines du bateau émettant une perturbation peut être calculée grâce aux puces GPS placées sur les narvals. Ainsi, nous pouvons modéliser l'effet de l'exposition des perturbations humaines sur l'émission de buzz des narvals et donc indirectement sur leur alimentation.

1.1 Modélisation de l'effet de l'exposition sur le taux d'émission

1.1.1 Processus de Poisson

Les données sont collectées toutes les secondes, ainsi nous disposons des temps de collecte T_j et $T_{j+1} = T_j + \Delta$ avec $T_0 = 0$ et $\Delta = 1$ seconde. On peut considérer $N(t)$, le nombre de buzz émis entre le début de la collecte et un instant t , comme un processus stochastique pour lequel $N(0) = 0$ et donnant $N(b) - N(a)$ le nombre de buzz émis dans l'intervalle $[a, b]$.

Les processus de comptage sont classiquement représentés par un processus de Poisson dont les accroissements sont indépendants et suivent une loi de Poisson : $N(t + \Delta) - N(t) \sim \text{Pois}$. Dans le cas où l'intensité λ du processus dépend de t , on parle de processus non-homogène et :

$$N(t + \Delta) - N(t) \sim \text{Pois}\left(\int_t^{t+\Delta} \lambda(u)du\right)$$

On a donc que :

$$\mathbb{P}(N(t + \Delta) - N(t) = k) = e^{-\int_t^{t+\Delta} \lambda(u)du} \frac{(\int_t^{t+\Delta} \lambda(u)du)^k}{k!}$$

En utilisant le développement limité de l'exponentielle et en considérant $\lambda(t)$ constante sur $[t, t + \Delta]$ on obtient :

$$\begin{aligned} \mathbb{P}(N(t + \Delta) - N(t) = 0) &= 1 - \lambda(t)\Delta + o(\Delta) \\ \mathbb{P}(N(t + \Delta) - N(t) = 1) &= \lambda(t)\Delta + o(\Delta) \\ \mathbb{P}(N(t + \Delta) - N(t) \geq 2) &= o(\Delta) \end{aligned}$$

Ainsi, si Δ est suffisamment petit (de l'ordre de la seconde par exemple), $o(\Delta)$ est négligeable et le processus $Y(t) := N(t + \Delta) - N(t)$ prend uniquement comme valeurs 0 ou 1. On peut alors ramener notre processus de comptage à un processus de Bernoulli : $Y(t) \sim \mathcal{B}(\lambda(t)\Delta)$, ce qui implique :

$$\mathbb{E}(Y(t)) = \mathbb{P}(Y(t) = 1) = \lambda(t)\Delta \underset{\Delta=1}{=} \lambda(t)$$

1.1.2 Modèle linéaire généralisé

Nous cherchons à estimer le lien entre l'intensité d'émission de buzz et plusieurs covariables, dont le niveau d'exposition aux perturbations. Nous avons vu dans la section précédente que les variables aléatoires $Y(t)$ de notre processus de comptage sont à valeurs dans 0, 1 et que $\mathbb{P}(Y(t) = 1) = \lambda(t)$. En se plaçant dans le cadre des modèles linéaires généralisés (GLM) il est donc possible d'exprimer la probabilité $\lambda(t)$ que $Y(t) = 1$ selon un vecteur de covariables $Z(t)$ (détailées dans la suite de cette section) en utilisant le logarithme comme fonction de lien :

$$\log(\lambda(t)) = \beta_0 + Z(t)^T \beta_Z$$

avec β_Z le vecteur de paramètres associés au vecteur $Z(t)$.

1.1.2.1 Effet de la profondeur

Pour se nourrir, les narvals doivent plonger profondément (plusieurs centaines de mètres), alors que le reste du temps elles restent “proches” (quelques dizaines de mètres) de la surface. Il faut donc inclure au modèle la covariable de profondeur à laquelle se trouvent les baleines quand elles émettent ou non des buzz. La relation entre l’émission de buzz et la profondeur n’étant pas linéaire, la profondeur a été remplacée par une spline cubique naturelle ayant pour noeuds les quantiles 1/3 et 2/3.

1.1.2.2 Effet de l’exposition

Le niveau d’exposition aux perturbations est représenté par l’inverse de la distance séparant la baleine du bateau quand un coup de feu est tiré. De même que pour la profondeur, la non-linéarité de la relation entre le niveau d’exposition et le taux d’émission de buzz est représentée par l’utilisation d’une spline cubique naturelle dont les noeuds sont les quantiles 1/3 et 2/3 des niveaux d’exposition.

1.1.3 Non-indépendance des observations

L’utilisation d’un processus de Poisson pour modéliser nos données de comptage implique l’indépendance des observations $Y(t)$. En pratique ce n’est pas le cas, aussi il faut modifier le modèle défini précédemment pour tenter de compenser cette absence d’indépendance.

1.1.3.1 Utilisation de modèles mixtes

Les données que nous utilisons correspondent à plusieurs individus, nous avons donc plusieurs observations par individu et celles-ci ne sont pas indépendantes. Pour palier ce défaut de modélisation et tenir compte de la spécificité des individus, nous utilisons des modèles mixtes en ajoutant un effet aléatoire b_i sur l’ordonnée à l’origine :

$$\log(\lambda_i(t)) = \beta_0 + b_i + \text{spline}(D_i(t))\beta_{D_{1:3}} + \text{spline}(E_i(t))\beta_{E_{1:3}}$$

où i dénote l’individu i et $b_i \sim \mathcal{N}(0, \sigma^2)$ l’effet aléatoire sur cet l’individu, σ^2 étant la variance inter-individuelle.

1.1.3.2 Caractère autorégressif du processus

De plus, l’émission d’un buzz à un instant t est corrélé à l’émission ou non de buzz aux instants précédents ; cet effet mémoire doit donc être intégré au modèle pour tenir compte de la dépendance des $Y_i(t)$. Pour cela nous introduisons K variables binaires d’autorégression codant l’émission d’un buzz aux instants $t - k$, $k \in \{1, \dots, K\}$. Le modèle résultant s’écrit alors :

$$\log(\lambda_i(t)) = \beta_0 + b_i + \text{spline}(D_i(t))\beta_{D_{1:3}} + \sum_{k=1}^K \alpha_k Y_i(t-k) + \text{spline}(E_i(t))\beta_{E_{1:3}}$$

Cette approche demande de fixer une mémoire maximale, et ainsi la valeur de K . Pour choisir la mémoire maximale optimale, nous avons fait varier K et utilisé le BIC comme mesure de la qualité des différents modèles correspondants sans inclure l'effet de l'exposition :

$$\log(\lambda_i(t)) = \beta_{D_0} + b_i + \text{spline}(D_i(t))\beta_{D_{1:3}} + \sum_{k=1}^K \alpha_k Y_i(t-k) \quad (1.1)$$

Nous choisissons la mémoire maximale K_{opt} du modèle minimisant ce critère. Pour éviter de parcourir tout l'ensemble $\{K_{min}, \dots, K_{max}\}$, nous avons utilisé la démarche proposée par l'Algorithme 1 permettant de restreindre l'ensemble de recherche au fur et à mesure que l'on s'approche de K_{opt} .

Algorithme 1 Réduction progressive de l'ensemble de recherche de K_{opt} .

```

 $from_k \leftarrow K_{min}; to_k \leftarrow K_{max}$                                 ▷ Bornes de l'intervalle de recherche
 $M \leftarrow 10$                                                                ▷ Nombre d'éléments évalués dans l'intervalle
while ( $to_k - from_k > 2$ ) do ▷ On s'arrête quand on évalué l'ensemble du voisinage du minimum
     $K_{1:M} \leftarrow NA$ 
     $BIC_{1:M} \leftarrow NA$ 
    for  $i \in \{1, \dots, M\}$  do
         $K_i \leftarrow from_k + (i - 1) * \lfloor \frac{to_k - from_k}{M-1} \rfloor$           ▷ Découpage en  $M$  éléments équidistants
         $\mathcal{M} \leftarrow$  ajustement du modèle de l'Equation (1.1) avec  $K_i$  éléments mémoire
         $BIC_i \leftarrow$  calcul du BIC de  $\mathcal{M}$ 
    end for
     $i_{opt} \leftarrow argmin BIC$ 
     $K_{opt} \leftarrow K_{i_{opt}}$ 
     $from_k \leftarrow K_{i_{opt}-1}; to_k \leftarrow K_{i_{opt}+1}$           ▷ Mise à jour des bornes en encadrant le minimum
end while

```

Le nombre de composants autorégressifs pouvant être grand, il faut les lier aux α_k avec un modèle de régression. Pour cela, nous avons utilisé une régression double bi-exponentielle :

$$BiExp(lag) = A_1 e^{-e^{lrc_1} lag} + A_2 e^{-e^{lrc_2} lag}$$

Cela permet de réduire le nombre de coefficients de K_{opt} à 4, ce qui est doublement bénéfique : le temps d'ajustement des modèles est grandement réduit et lors de la construction des intervalles de confiance de nos coefficients, l'accumulation des variances est limitée.

Le modèle complet s'exprime donc ainsi :

$$\log(\lambda_i(t)) = \beta_0 + b_i + \text{spline}(D_i(t))\beta_{D_{1:3}} + \sum_{k=1}^{K_{opt}} (A_1 e^{-e^{lrc_1} k} + A_2 e^{-e^{lrc_2} k}) Y_i(t-k) + \text{spline}(E_i(t))\beta_{E_{1:3}}$$

1.1.4 Effet de médiation de la profondeur

Il est possible que l'exposition à des perturbations conduisent les narvals à :

1. émettre moins de buzz,

2. moins plonger ou plonger moins profondément.

Mais il existe également un lien entre la profondeur d'immersion des baleines et leur production de buzz. Aussi, il se peut que l'exposition ait un lien direct sur l'émission de buzz et un lien indirect via son effet sur la profondeur.

Afin de représenter uniquement le lien direct, les coefficients des covariables autres que l'exposition sont estimés sans inclure celle-ci à partir des observations effectuées sans soumettre les animaux à des perturbations :

$$\log(\lambda_i(t)) = \beta_{D_0} + b_i + \text{spline}(D_i(t))\beta_{D_{1:3}} + \sum_{k=1}^{K_{opt}} A_1 e^{-e^{lrc_1} k} + A_2 e^{-e^{lrc_2} k} Y_i(t-k) \quad (1.2)$$

Nous obtenons donc les estimations $\widehat{\beta_D}$, $\widehat{A_{1:2}}$, $\widehat{lrc_{1:2}}$ qui sont ensuite injectées dans le modèle complet au moyen d'un terme d'"offset" (non réestimé) :

$$offset_i(t) = \widehat{\beta_{D_0}} + \text{spline}(D_i(t))\widehat{\beta_{D_{1:3}}} + \sum_{k=1}^{K_{opt}} (\widehat{A_1} e^{-e^{\widehat{lrc_1}} k} + \widehat{A_2} e^{-e^{\widehat{lrc_2}} k}) Y_i(t-k) \quad (1.3)$$

Le modèle incluant l'exposition est donc reformulé ainsi :

$$\log(\lambda_i(t)) = \beta_{E_0} + b_i + offset_i(t) + \text{spline}(E_i(t))\beta_{E_{1:3}} \quad (1.4)$$

Et les paramètres estimés $\widehat{\beta_E}$ permettront d'évaluer l'effet de l'exposition par rapport à des conditions "normales".

1.1.5 Intervalles de confiance

1.1.5.1 Construction par approche Monte-Carlo

Après avoir estimé les coefficients associés à l'exposition aux perturbations, nous souhaitons construire les intervalles de confiance de ces estimations. L'approche classique de calcul des intervalles de confiance se basant sur la seule variance estimée des coefficients d'exposition donnerait ici des résultats incorrects, la variance des coefficients de profondeur et d'autorégression ne serait alors pas prise en compte car tuée par l'utilisation de l'offset.

1.1.5.1.1 Utilisation de lois normales univariées Dans un premier temps, nous avons donc ajusté le modèle sans exposition décrit par l'Equation (1.2) afin d'obtenir la moyenne et la variance empiriques de ses coefficients. Nous avons ensuite répété le tirage des coefficients sans exposition selon 8 lois normales univariées paramétrées par leurs statistiques empiriques ; calculé le terme d'offset selon l'Equation (1.3) à partir des réalisations obtenues ; et estimé les coefficients d'exposition $\widehat{\beta_E^k}$ du modèle correspondant à l'Equation (1.4). Les coefficients $\widehat{\beta_E^k}$ estimés à chaque itération constituent ainsi un échantillon dont nous utilisons les quantiles empiriques $\alpha/2$ et $1 - \alpha/2$ comme bornes de l'intervalle de confiance au niveau α de chacun des coefficients d'exposition.

1.1.5.1.2 Utilisation de lois normales multivariées Le tirage selon des lois normales univariées implique que nous ne considérons pas la covariance existante entre les estimations des coefficients $\widehat{\beta}_D$, $\widehat{A}_{1:2}$ et $\widehat{lrc}_{1:2}$. Afin d'y remédier nous avons employé un tirage suivant 2 lois normales multivariées de dimension 4 (une pour les coefficients $\widehat{\beta}_D$ et une pour $\widehat{A}_{1:2}$ et $\widehat{lrc}_{1:2}$, dont le vecteur de moyenne est toujours constitué des moyennes empiriques des coefficients sans exposition, mais dont les matrices de variance-covariance ne sont pas diagonales).

1.1.5.1.3 Variance-covariance des coefficients autorégressifs La matrice de variance-covariance permettant de tirer les coefficients $\widehat{A}_{1:2}$ et $\widehat{lrc}_{1:2}$ était construite à partir des estimations de la régression double bi-exponentielle, ce qui signifie que nous ne captions pas directement la variabilité du phénomène autorégressif mais plutôt celle de la régression double bi-exponentielle. Pour palier cette approximation, nous avons recalculé la matrice de variance-covariance des coefficients $\widehat{A}_{1:2}$ et $\widehat{lrc}_{1:2}$ par une autre procédure Monte-Carlo en répétant l'ajustement de la régression double bi-exponentielle pour des coefficients mémoire tirés selon une loi normale multivariée. Cette approche est décrite par l'Algorithme 2.

Algorithme 2 Procédure Monte-Carlo d'obtention des intervalles de confiance.

```

 $\mathcal{M}_0 \leftarrow$  ajustement du modèle correspondant à l'Equation (1.1)
 $\mu_D, \Sigma_D, \mu_{AR}, \Sigma_{AR} \leftarrow$  estimation à partir de  $\mathcal{M}_0$ 
 $K \leftarrow 1000$  ▷ Nombre de répétitions des procédures
▷ 1ère procédure

 $\widehat{\beta}_{A_{1:2}}^{(1:K)} \leftarrow NA$ 
 $\widehat{\beta}_{lrc_{1:2}}^{(1:K)} \leftarrow NA$ 
for  $k \in \{1, \dots, K\}$  do
     $\hat{\alpha} \leftarrow \mathcal{N}(\mu_{AR}, \Sigma_{AR})$  ▷ Tirage des coefficients autorégressifs
     $\mathcal{M}_{exp} \leftarrow$  ajustement de la régression double bi-exponentielle utilisant  $\hat{\alpha}$ 
     $\widehat{\beta}_{A_{1:2}}^{(k)}, \widehat{\beta}_{lrc_{1:2}}^{(k)} \leftarrow$  coefficients de  $\mathcal{M}_{exp}$ 
end for
 $\mu_{exp}, \Sigma_{exp} \leftarrow$  calculés à partir de  $\widehat{\beta}_{A_{1:2}}, \widehat{\beta}_{lrc_{1:2}}$  ▷ 2ème procédure

 $\widehat{\beta}_E^{(1:K)} \leftarrow NA$ 
for  $k \in \{1, \dots, K\}$  do
     $\widehat{\beta}_D \leftarrow \mathcal{N}(\mu_D, \Sigma_D)$  ▷ Tirage des coefficients de profondeur
     $\widehat{A}_{1:2}, \widehat{lrc}_{1:2} \leftarrow \mathcal{N}(\mu_{exp}, \Sigma_{exp})$  ▷ Tirage des coefficients de la double bi-exponentielle
    calcul de l'offset selon l'Equation (1.3) avec  $\widehat{\beta}_D, \widehat{A}_{1:2}, \widehat{lrc}_{1:2}$ 
     $\mathcal{M} \leftarrow$  ajustement du modèle donné par l'Equation (1.4) en fixant l'offset
     $\widehat{\beta}_E^{(k)} \leftarrow$  coefficients de profondeur de  $\mathcal{M}$ 
end for
 $IC_\alpha \leftarrow [q_{\alpha/2}^{\widehat{\beta}_E}, q_{1-\alpha/2}^{\widehat{\beta}_E}]$  ▷ Estimation via les quantiles empiriques

```

1.1.5.1.4 Sans passer par la régression double bi-exponentielle L'intérêt de l'utilisation de la régression double bi-exponentielle est de :

1. réduire le temps d'ajustement des modèles linéaires,
2. éviter d'accumuler les variances des 60 coefficients de mémoire.

En fixant les coefficients autorégressifs, avoir 4 ou 60 coefficients pour la mémoire n'importe plus ; et en utilisant une loi normale multivariée, nous devrions également ne plus accumuler directement les variances des coefficients. Nous pouvons donc envisager de nous passer de la régression double bi-exponentielle, et de tirer les 60 coefficients mémoire et les 4 coefficients de profondeur directement dans une seule loi normale multivariée dont les paramètres sont obtenus après ajustement du modèle sans exposition de l'Equation (1.1). Ainsi nous revenons à une seule procédure Monte-Carlo (la deuxième dans l'Algorithme 2) et l'offset utilisé dans l'Equation (1.4) du modèle complet devient :

$$offset_i(t) = \widehat{\beta_{D_0}} + spline(D_i(t))\widehat{\beta_{D_{1:3}}} + \sum_{k=1}^{K_{opt}} \widehat{\alpha_k} Y_i(t-k)$$

1.1.5.2 Estimation de bandes de prédiction via la méthode Delta

Une fois les coefficients d'exposition $\widehat{\beta_E}$ estimés nous pouvons prédire le taux d'émission de buzz sur l'ensemble d'un intervalle de niveau d'exposition aux perturbations. Afin d'associer une bande de confiance à la prédiction moyenne nous avons appliqué la méthode Delta pour calculer la variance du taux d'émission.

Tout d'abord, d'après le Théorème Central Limite on a : $\sqrt{n}(\widehat{\beta} - \beta) \sim \mathcal{N}(0, \Sigma)$ avec $\widehat{\beta}$ les coefficients estimés du modèle et Σ leur variance.

Etant donné que nous utilisons un GLM avec un lien log, le taux d'émission est lié aux coefficients du modèle via la relation $\lambda(t) = \exp(X(t)\beta) := f(\beta)$. Le gradient de la fonction f vaut $\nabla f(\beta) = Xf(\beta)$.

Ainsi, quand on applique la méthode Delta on obtient :

$$\begin{aligned} \sqrt{n}(f(\widehat{\beta}) - f(\beta)) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla f(\beta)^T \Sigma \nabla f(\beta)) \\ &\Leftrightarrow \sqrt{n}(\widehat{\lambda(t)}) - \lambda(t) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda(t)^2 X(t)^T \Sigma X(t)) \end{aligned}$$

donc

$$\frac{\sqrt{n}}{\widehat{\lambda(t)} \sqrt{X(t)^T \widehat{\Sigma} X(t)}} (\widehat{\lambda(t)}) - \lambda(t) \rightsquigarrow \mathcal{N}(0, 1)$$

On en déduit la bande de confiance au niveau α :

$$IC_\alpha(\lambda(t)) = [\widehat{\lambda(t)} \pm q_{1-\alpha/2}^{\mathcal{N}} \frac{\widehat{\lambda(t)} \sqrt{X(t)^T \widehat{\Sigma} X(t)}}{\sqrt{n}}]$$

où $q_{1-\alpha/2}^{\mathcal{N}}$ est le quantile $1 - \alpha/2$ de la loi normale centrée réduite.

1.2 Résultats

1.2.1 Note sur le temps d'ajustement des modèles

Le temps d'ajustement des modèles, et en particulier des modèles mixtes, augmente fortement lorsque que le nombre de paramètres à ajuster augmente.

Comme nous pouvons le voir sur la Figure 1.1 cette augmentation est linéaire pour les modèles classiques, alors que pour les modèles mixtes celle-ci est quadratique.

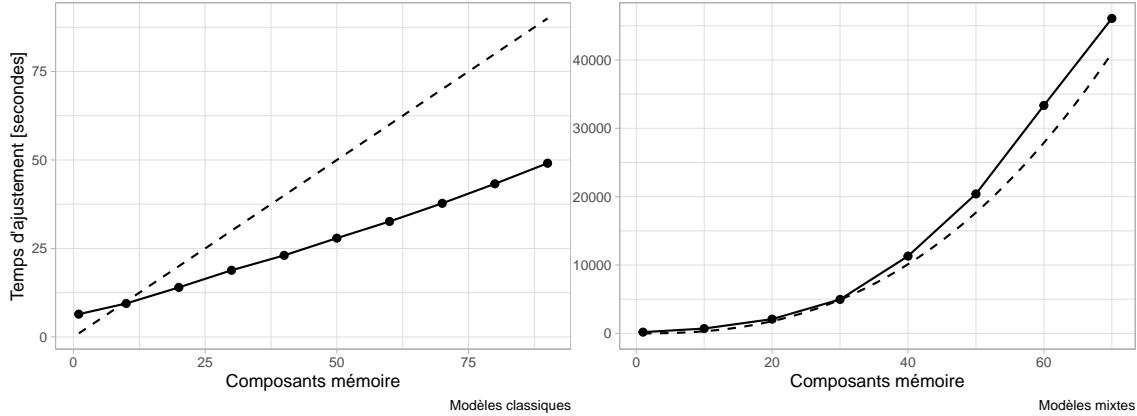


Figure 1.1: Courbes pleines : temps d'ajustement (en secondes) des modèles en fonction de la mémoire maximum ; courbes en pointillés : $f(x) = x$ à gauche et $f(x) = x^{2.5}$ à droite.

1.2.2 Recherche de la mémoire optimale

Le temps d'ajustement des modèles mixtes étant nettement plus important que ceux des modèles sans effets aléatoires, nous avons dans un premier temps exclu ces effets du modèle sans exposition pour estimer la mémoire optimale. Nous avons choisi pour la recherche $K_{min} = 1$, $K_{max} = 300$, $M = 10$. Nous obtenons une mémoire optimale de 60 secondes. Cela nous a permis de restreindre tout de suite l'ensemble de recherche initial à $K_{min} = 1$, $K_{max} = 300$ quand nous avons considéré le modèle incluant les effets aléatoires. De même que précédemment, la mémoire optimale est égale à 60 secondes. La Figure 1.2 permet de voir que dans les deux cas les optimums semblent bien correspondre à des minimums globaux.

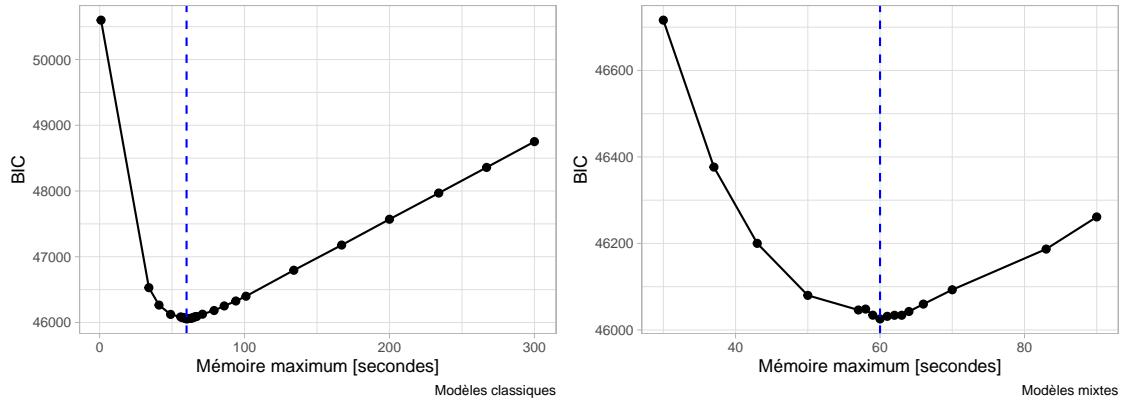


Figure 1.2: BIC en fonction de la mémoire maximum.

1.2.3 Régression double bi-exponentielle sur les coefficients autorégressifs

La Figure 1.3 permet de comparer la régression double bi-exponentielle et les composantes de la mémoire ajustées pour un décalage maximum de 60. Nous pouvons constater que l'ajustement par la double bi-exponentielle est très fidèle aux 60 coefficients initiaux.

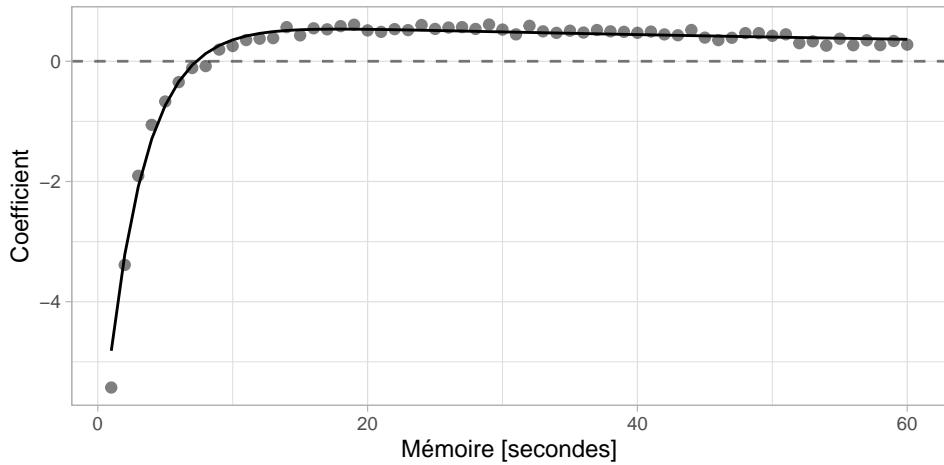


Figure 1.3: Régression double bi-exponentielle des coefficients autorégressifs.

Les 4 coefficients obtenus suite à la régression double bi-exponentielle sont présentés dans la Table 1.1. Nous pouvons noter que leurs erreurs standard sont faibles, ce qui est cohérent avec la validation visuelle de la régression.

Table 1.1: Coefficients autorégressifs obtenus par régression double bi-exponentielle

	moyenne	erreur standard
A_1	-7.741	0.329
lrc_1	-1.055	0.041
A_2	0.654	0.043
lrc_2	-4.633	0.188

1.2.4 Impact de l'exposition aux perturbations sur le taux d'émission de buzz

La Table 1.2 expose les coefficients associés à l'exposition du modèle mixte. Interpréter les valeurs des coefficients liés à des splines étant peu pertinent, nous préférons nous référer à la lecture de la Figure 1.4 sur laquelle nous représentons les estimations du modèle pour le taux d'émission de buzz selon la distance avec le bateau. Il apparaît alors clairement que plus le bateau est loin, plus le taux d'émission de buzz est élevé.

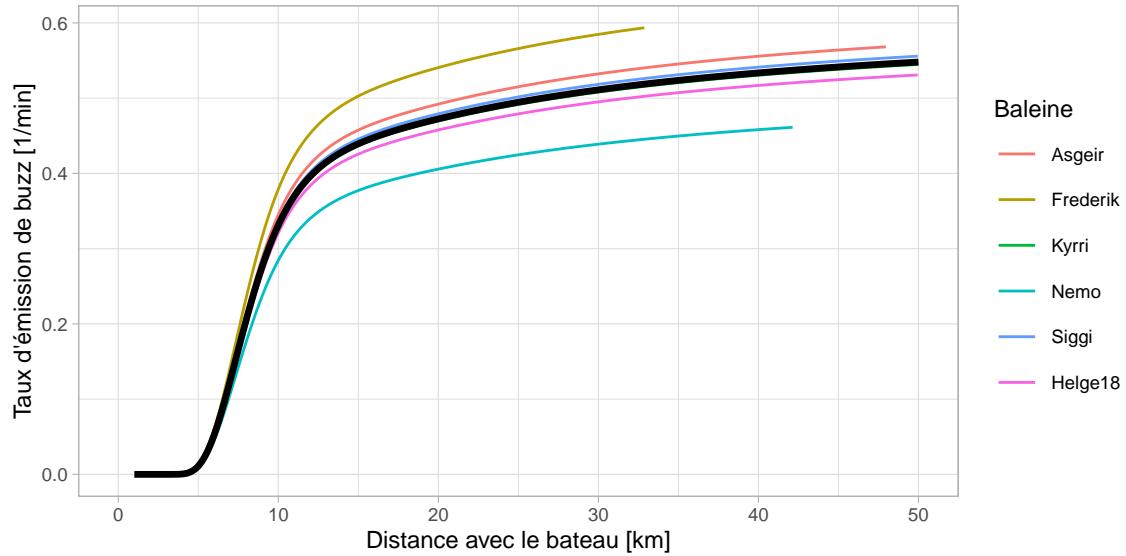


Figure 1.4: Evolution du taux d'émission de buzz selon la distance au bateau.

Table 1.2: Coefficients d'exposition.

	moyenne	erreur standard
β_{E_1}	-1.146	0.002
β_{E_2}	-58.872	0.054
β_{E_3}	-112.790	0.107

Nous voulons vérifier que notre processus de comptage est bien représenté par un processus de Poisson et que nous avons bien tenu compte de la corrélation entre ses accroissements. Pour cela, nous nous sommes intéressé aux résidus uniformes du modèle.

Pour vérifier l'absence de corrélation, nous avons représenté sur la Figure 1.5 l'autocorrélation des résidus en fonction du décalage temporel, ainsi que les résidus en fonction des résidus précédents. Nous remarquons qu'aucun motif de corrélation ne semble apparaître, ce qui vient valider notre approche pour inclure au modèle le lien naturel existant entre les émissions de buzz.

Le Q-Q plot de la Figure 1.5 confirme lui que le processus de Poisson est bien adapté aux données de comptage des buzz étant donné que les quantiles des résidus correspondent à ceux de la loi uniforme.

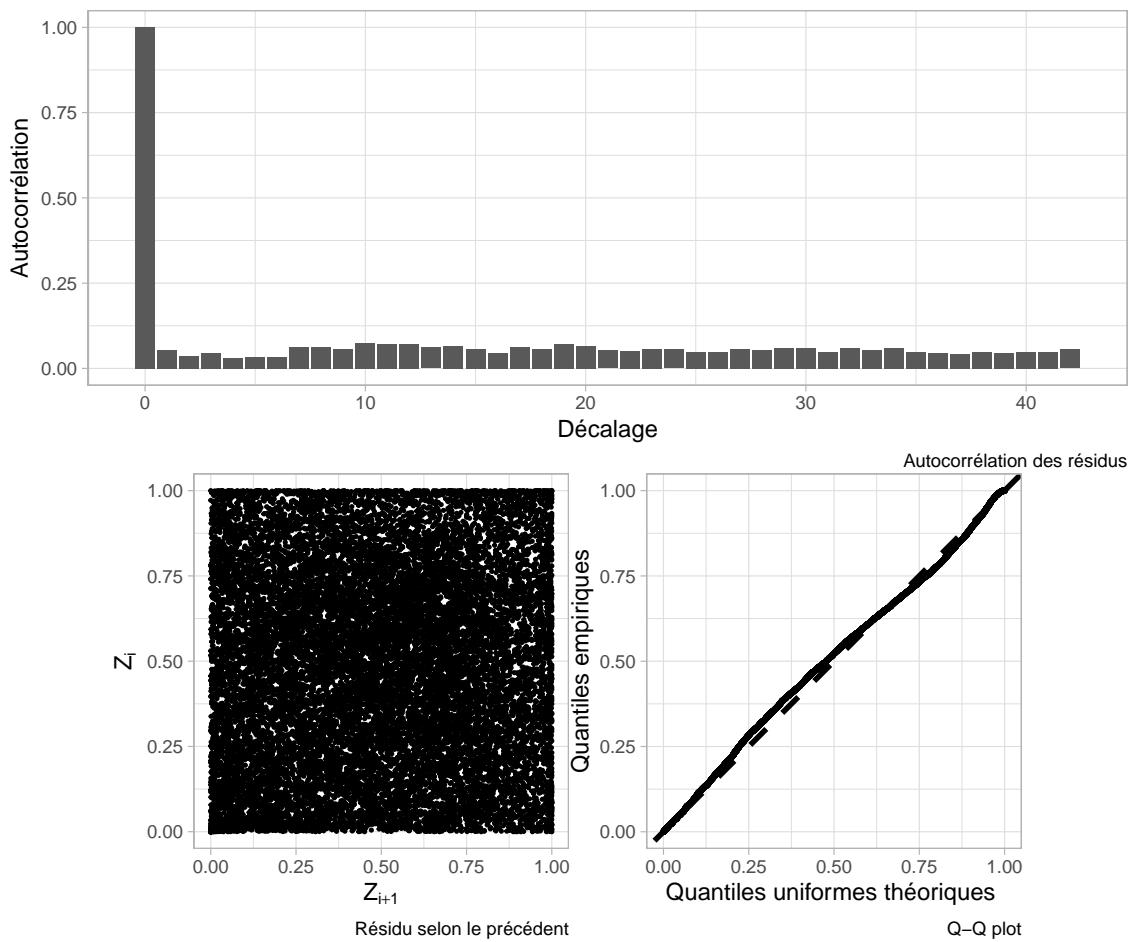


Figure 1.5: Validation graphique du modèle.

1.2.5 Intervalles de confiance

1.2.5.1 Coefficients d'exposition

1.2.5.1.1 Utilisation de lois normales univariées Comme détaillé dans la section 1.1.5.1.1, nous avons commencé par considérer que les coefficients autorégressifs et de profondeur suivaient chacun une loi normale centrée sur leur estimation moyenne et avec une variance égale au carré de leur erreur standard. Nous avons donc 4 lois normales univariées pour les coefficients de la régression double bi-exponentielle et 4 autres pour la spline sur la profondeur.

Sur la Figure 1.6 nous avons représenté les courbes des fonctions double bi-exponentielles ainsi générées. Nous pouvons voir que leurs allures semblent toujours suivre celle de la régression initiale.

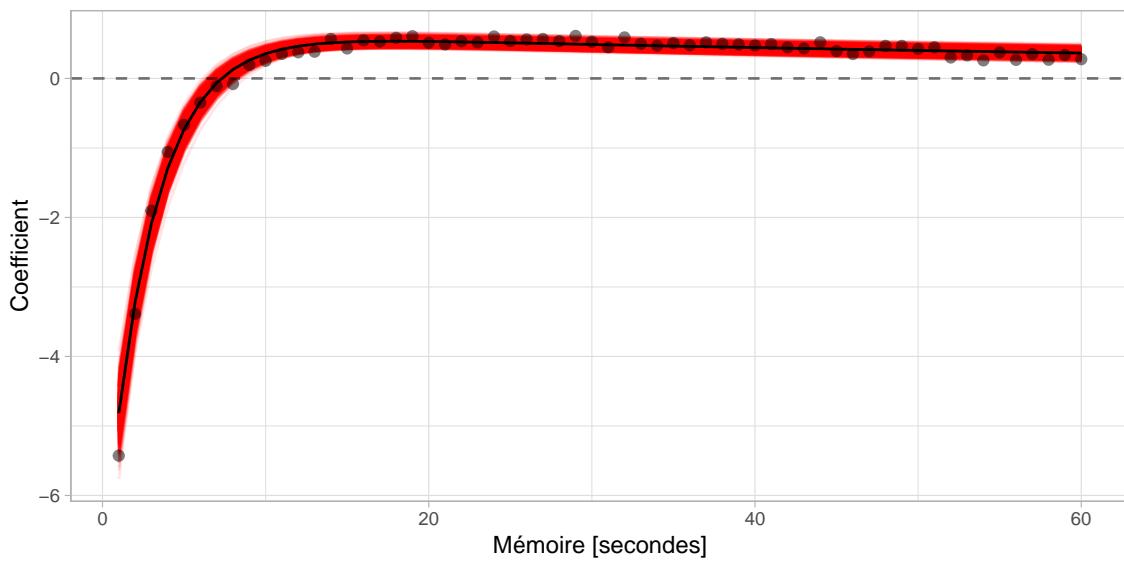


Figure 1.6: Variation des coefficients autorégressif selon des lois normales univariées.

Nous avons fait de même avec les coefficients de profondeur, et nous pouvons remarquer sur la Figure 1.7 que cette fois-ci certaines courbes s'éloignent sensiblement de l'interpolation moyenne.

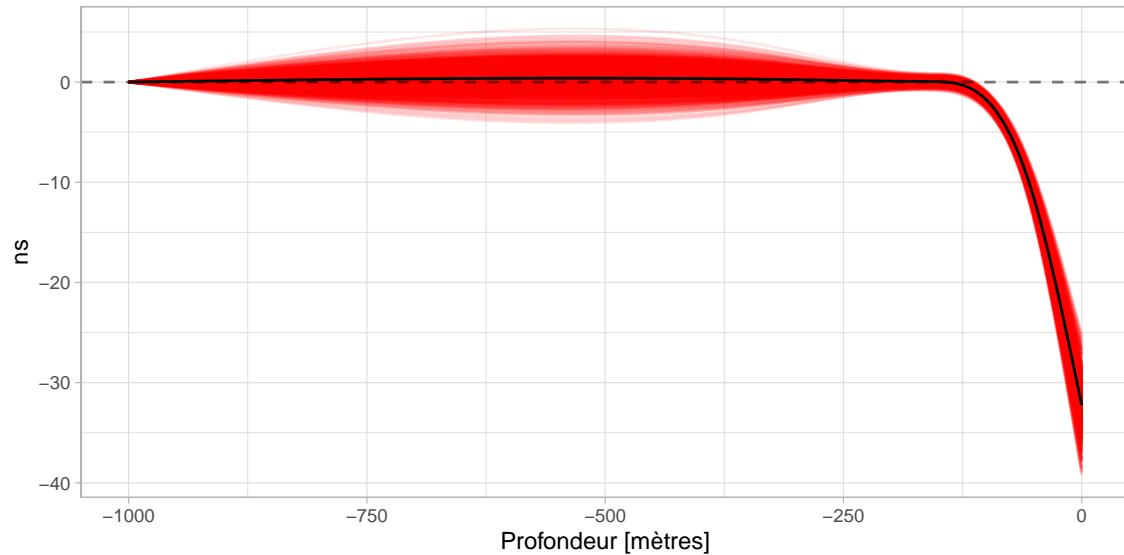


Figure 1.7: Variation des coefficients de profondeur selon des lois normales univariées.

Les intervalles de confiance calculés avec la procédure Monte-Carlo sont donnés sur la Table 1.3. Nous avons également affiché les valeurs médianes des intervalles de confiance calculés à chaque répétition sur la base de la variation des coefficients d'exposition uniquement. Il est flagrant que les intervalles Monte-Carlo sont bien plus larges et ne permettent en aucun cas de conclure sur un effet de l'exposition sur le taux d'émission de buzz.

Table 1.3: Intervalles de confiance dans le cas de normales univariées.

	Monte-Carlo		Erreur standard	
	inf	sup	inf	sup
β_0	-6.442	-2.968	-4.639	-4.494
β_{E_1}	-11.899	11.294	-1.430	-1.422
β_{E_2}	-152.959	1.048	-56.015	-55.803
β_{E_3}	-306.429	11.707	-106.847	-106.431

La Figure 1.8 donne une représentation graphique de ces intervalles (en rouge les “MC” et en bleu les “SE”), ainsi que des distributions des coefficients.

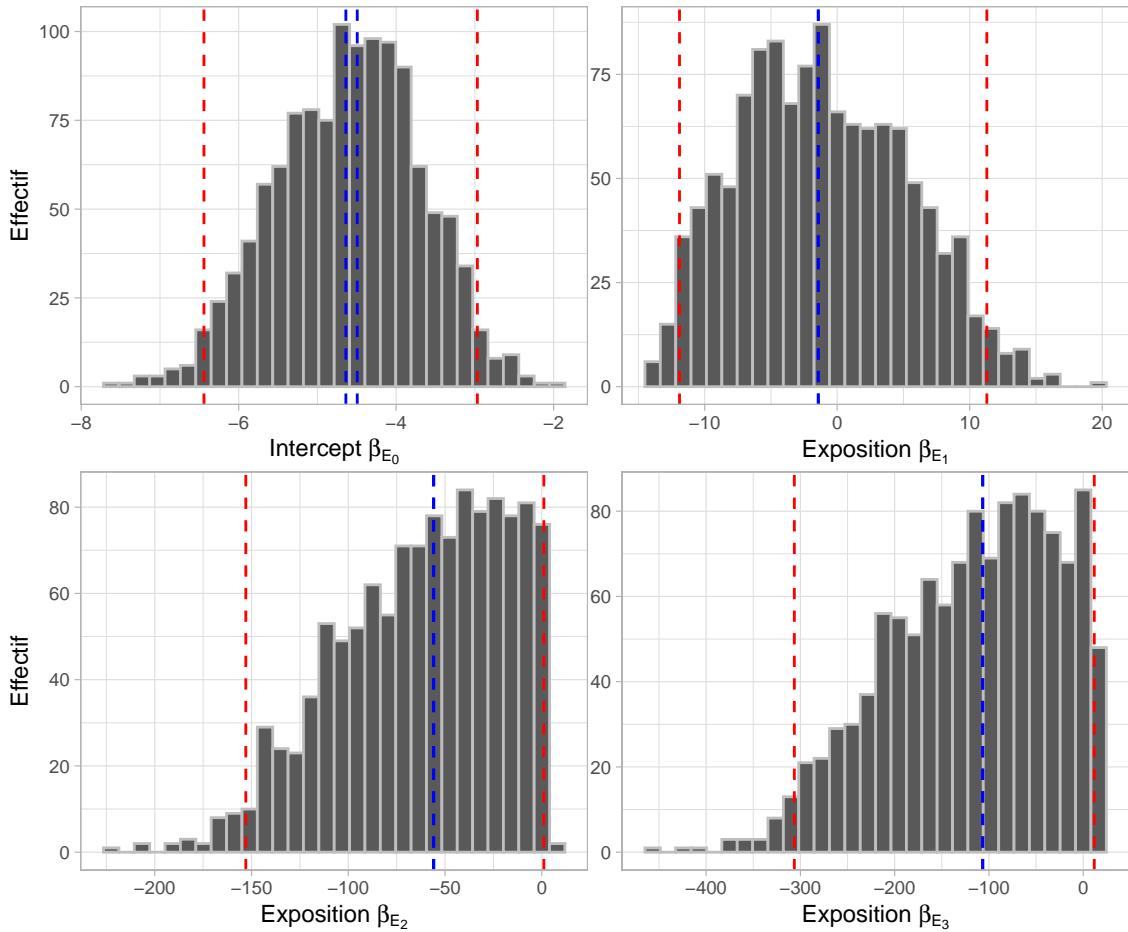


Figure 1.8: Intervalles de confiance dans le cas de normales univariées.

1.2.5.1.2 Utilisation de lois normales multivariées L'une des explications aux très larges intervalles de confiance observés dans la section précédente pourrait être que nous avons accumulé les variances des coefficients sans tenir compte des probables covariances existant entre les coefficients. Afin de corriger cela nous avons répété l'approche décrite précédemment, mais en tirant les coefficients dans 2 lois normales multivariées : une pour les coefficients de la régression double bi-exponentielle et une pour les coefficients de profondeur. Pour plus de détails sur l'obtention des paramètres des lois normales, il est possible de revenir à la section 1.1.5.1.2.

La Figure 1.9 montre que la régression double bi-exponentielle initiale est encore plus fidèlement suivie qu'auparavant.

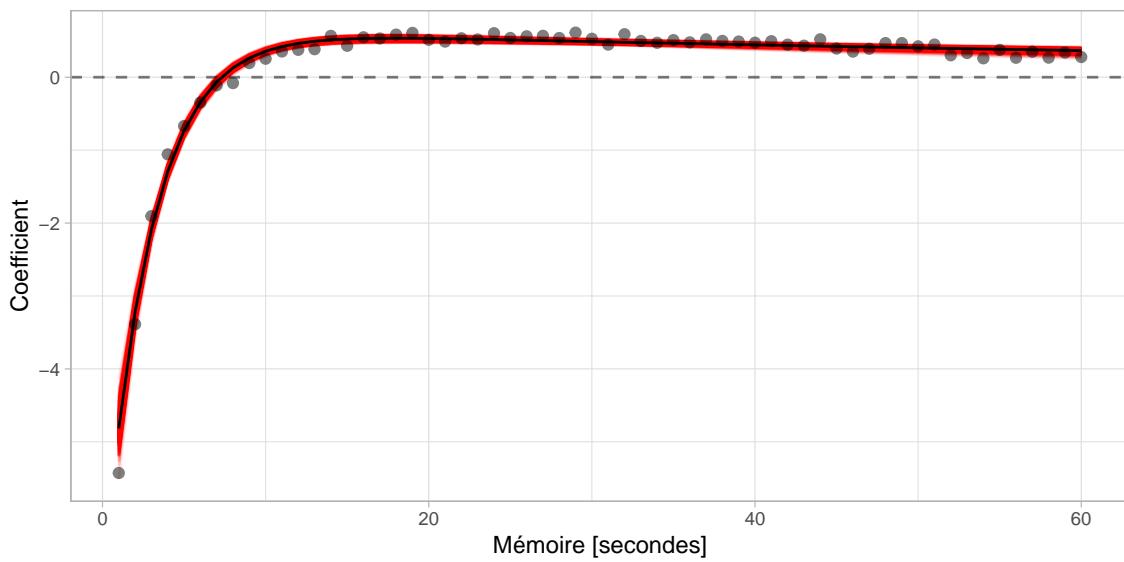


Figure 1.9: Variation des coefficients autorégressifs selon une loi normale multivariée.

Et surtout, comme nous pouvons le voir sur la Figure 1.10, il en va de même pour la profondeur, alors que précédemment les tirages donnaient des courbes fortement éloignées de celle attendue.

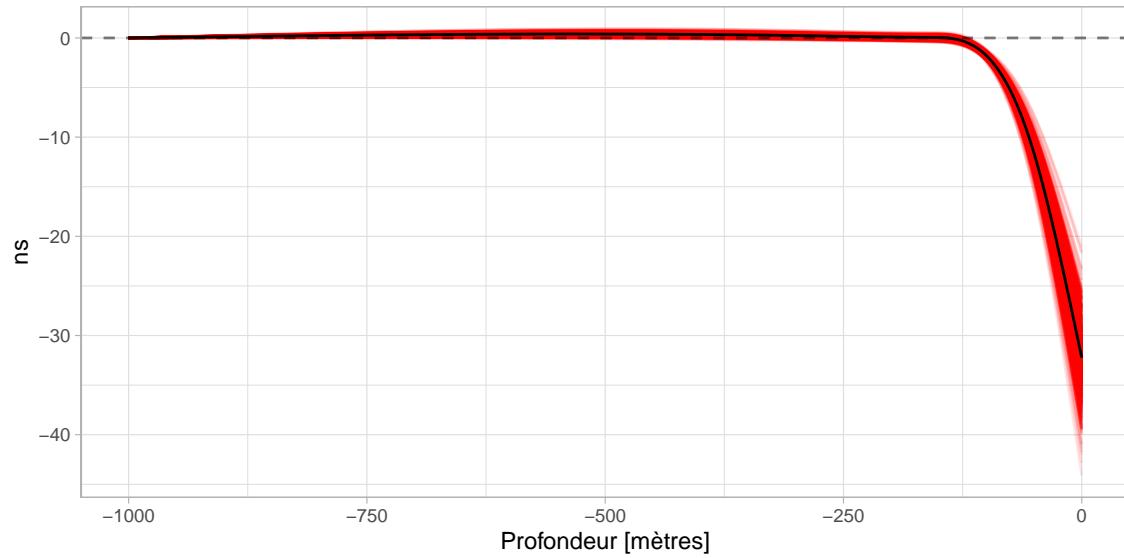


Figure 1.10: Variation des coefficients de profondeur selon une loi normale multivariée.

Sur la Table 1.4 nous pouvons constater que les intervalles de confiance estimés en utilisant des lois normales multivariées sont nettement plus petits et peuvent conduire à conclure sur un effet de l'exposition sur le taux d'émission de buzz.

Table 1.4: Intervalles de confiance dans le cas de normales multivariées.

	Monte-Carlo		Erreur standard	
	inf	sup	inf	sup
β_{E_0}	-4.888	-4.273	-4.632	-4.498
β_{E_1}	-1.971	-0.480	-1.237	-1.229
β_{E_2}	-64.615	-52.024	-58.310	-58.097
β_{E_3}	-124.493	-98.742	-111.628	-111.209

Nous pouvons également voir sur la Figure 1.11 que les distributions semblent normales, contrairement à celles observées avec les lois univariées.

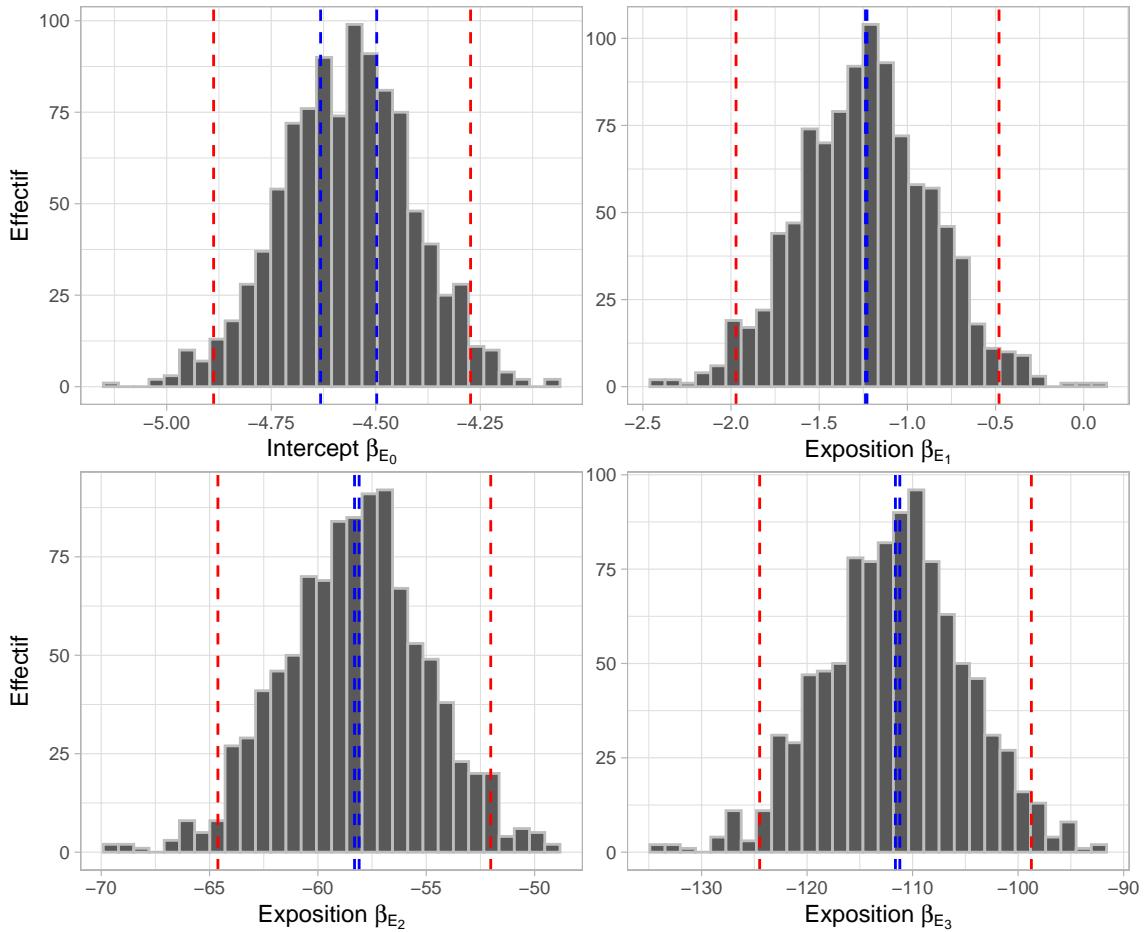


Figure 1.11: Intervalles de confiance dans le cas de normales multivariées.

1.2.5.1.3 Variance-covariance des coefficients autorégressifs Bien que plus intéressante, la procédure que nous avons mise en place capte la variabilité de la régression double bi-exponentielle et non celle du phénomène autorégressif en lui-même. Nous reprenons donc la procédure Monte-Carlo employée dans la section précédente, mais avec un vecteur de moyennes et une matrice de variance-covariance obtenus via une seconde procédure Monte-Carlo faisant varier les coefficients autorégressifs. L'ensemble de cette procédure est décrite par l'Algorithme 2 de la section 1.1.5.1.3.

Nous pouvons contrôler sur la Figure 1.12 que les régressions double bi-exponentielles obtenues restent cohérentes.

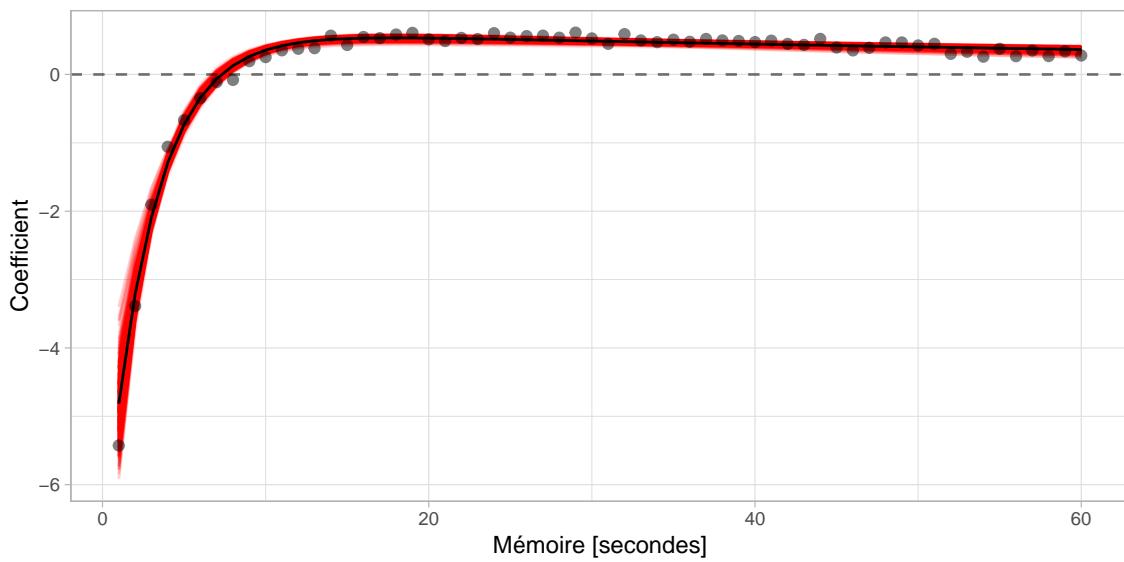


Figure 1.12: Variation des coefficients autorégressif selon une loi normale multivariée estimée par Monte-Carlo.

La Table 1.5 et la Figure 1.13 permettent de vérifier que bien que les intervalles calculés ainsi sont légèrement plus grands que les précédents, ils permettent toujours d'interpréter les coefficients d'exposition ajustés par le modèle.

Table 1.5: Intervalles de confiance dans le cas d'une normale multivariée estimée par Monte-Carlo.

	Monte-Carlo		Erreur standard	
	inf	sup	inf	sup
β_{E_0}	-4.866	-4.275	-4.645	-4.510
β_{E_1}	-1.981	-0.426	-1.122	-1.114
β_{E_2}	-65.140	-51.949	-59.343	-59.128
β_{E_3}	-125.572	-98.545	-113.695	-113.274

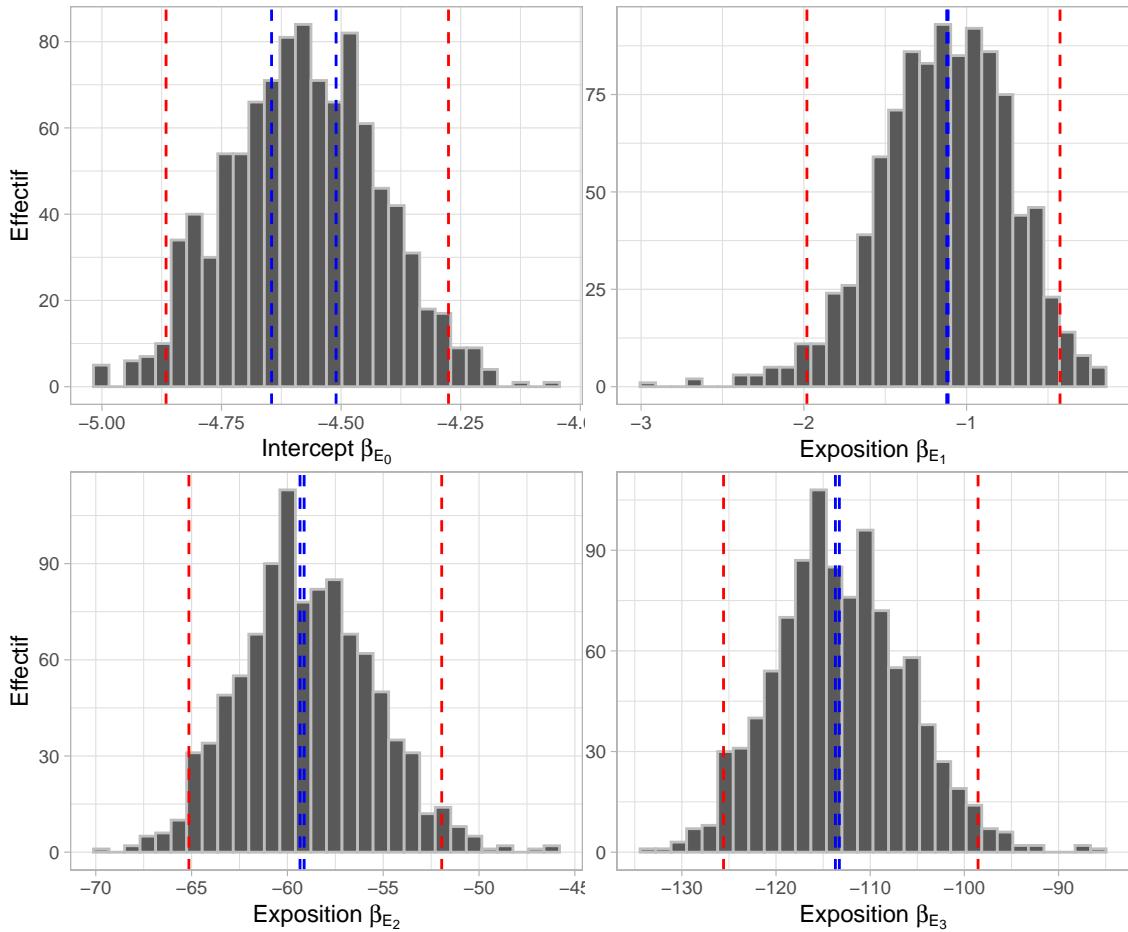


Figure 1.13: Intervalles de confiance dans le cas d'une normale multivariée estimée par Monte-Carlo.

1.2.5.1.4 Sans passer par la régression double bi-exponentielle Pour les raisons évoquées dans la section 1.1.5.1.4, il est également pertinent d'observer les intervalles de confiance des coefficients d'exposition en ayant calculé le terme d'offset directement à partir des coefficients autorégressifs, sans utiliser de régression double bi-exponentielle.

Nous pouvons voir sur la Table 1.6 et la Figure 1.14 qu'avec cette approche plus directe, les intervalles de confiance sont quasiment identiques à ceux obtenus dans la section précédente, et même plus petit pour l'ordonnée à l'origine.

Table 1.6: Intervalles de confiance dans le cas d'une loi normale multivariée sans double bi-exponentielle.

	Monte-Carlo		Erreur standard	
	inf	sup	inf	sup
β_{E_0}	-4.877	-4.261	-4.655	-4.513
β_{E_1}	-1.782	-0.566	-1.187	-1.178
β_{E_2}	-64.410	-52.093	-58.181	-57.969
β_{E_3}	-124.003	-99.056	-111.401	-110.984

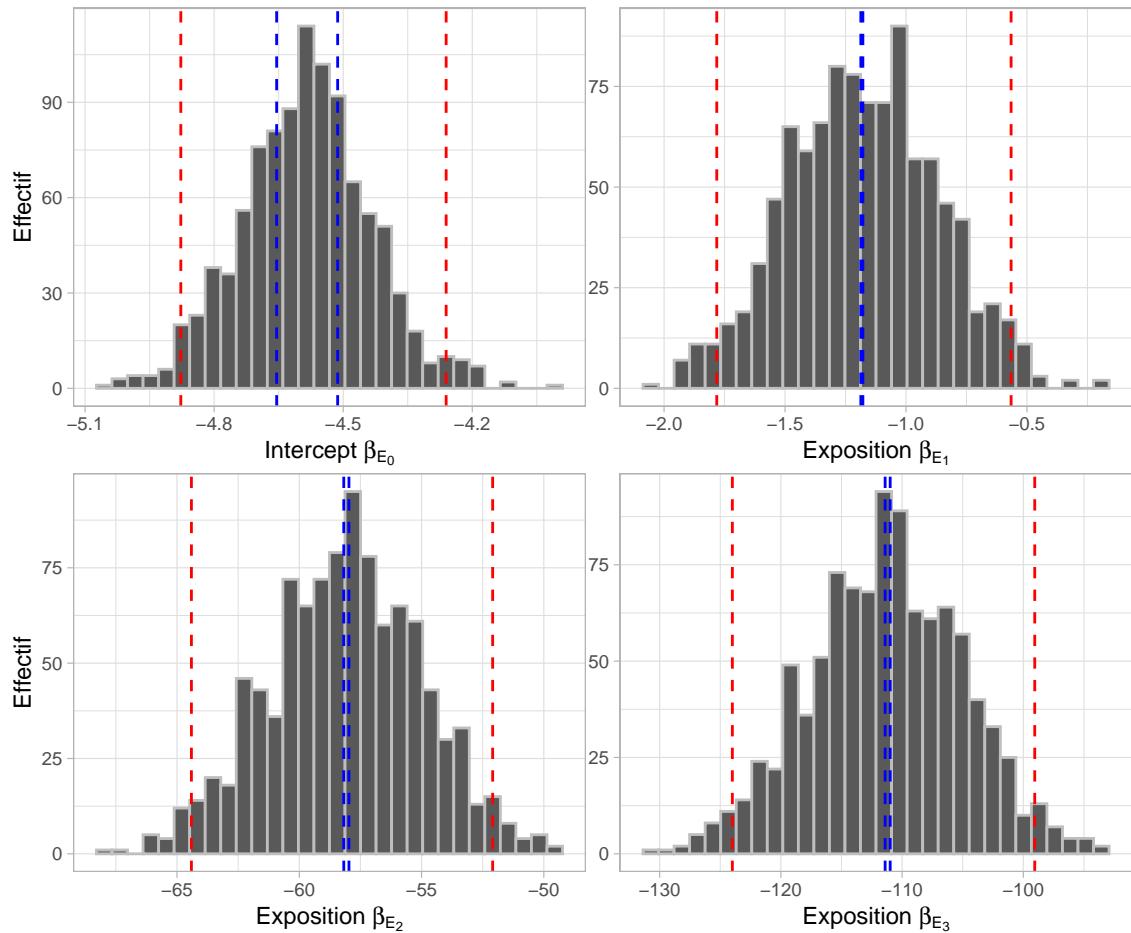


Figure 1.14: Intervalles de confiance dans le cas d'une loi normale multivariée sans double bi-exponentielle.

1.2.5.2 Pourcentage du taux normal d'émission de buzz

Il est rassurant de voir que les intervalles de confiance des coefficients d'exposition obtenus par Monte-Carlo restent raisonnablement petits, cependant comme évoqué précédemment, il n'est pas possible de lier une augmentation ou une diminution du taux d'émission de buzz en fonction de la distance avec le bateau à partir de leur valeur. C'est pourquoi nous avons représenté graphiquement cette évolution sur la Figure 1.4, mais cette visualisation ne proposait pas d'intervalle de confiance ou de bande de prédiction.

Nous avons employé la méthode Delta décrite dans la section 1.1.5.2 pour construire une bande de confiance autour de l'estimation du pourcentage du taux normal d'émission de buzz en fonction de la profondeur. Ce pourcentage est obtenu en faisant le rapport entre l'intensité estimée avec et sans perturbation. La Figure 1.15 permet de voir que la bande de confiance tracée en rouge est fine. Nous sommes donc confiants dans la lecture de la courbe qui illustre que soumis à des perturbations à une distance inférieure à 15 kilomètres, les narvals commencent à émettre nettement moins de buzz que dans des conditions normales. Afin de comparer l'approche de la méthode Delta, nous avons utilisé la fonction *predictInterval* du package **merTools**. Celle-ci permet d'obtenir une bande de prédiction pour des modèles mixtes par une approche Monte-Carlo. Comme nous pouvons le voir sur la Figure 1.15, la bande obtenue ainsi (tracée en bleue) suit la même forme que celle de la méthode Delta.

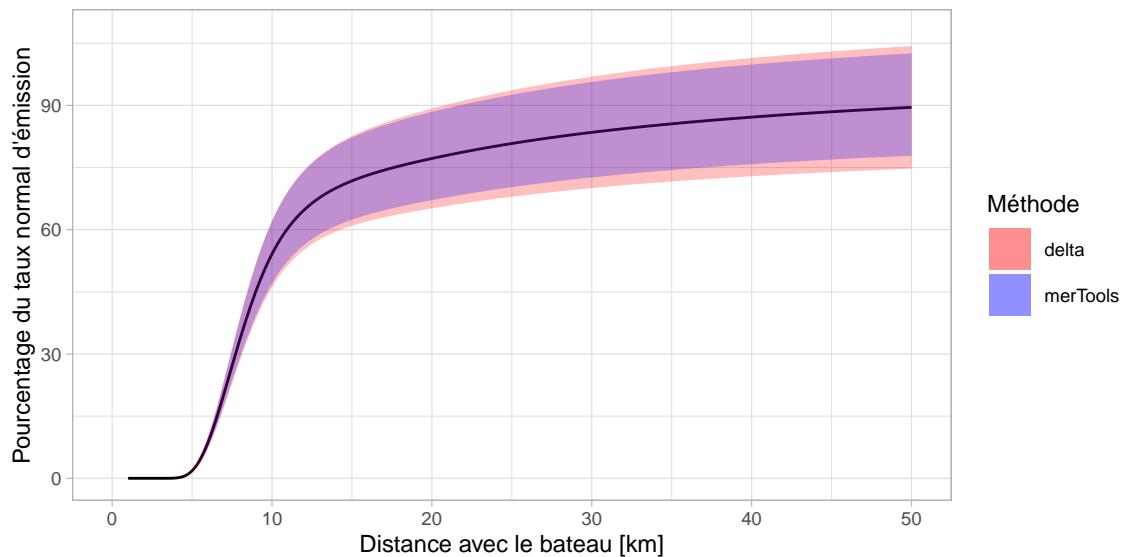


Figure 1.15: Pourcentage du taux normal d'émission de buzz selon la distance au bateau.

Références

- [1] Outi M. Tervo et al. “Narwhals react to ship noise and airgun pulses embedded in background noise”. In: *Biology Letters* 17.11 (2021), p. 20210220. DOI: [10.1098/rsbl.2021.0220](https://doi.org/10.1098/rsbl.2021.0220). eprint: <https://royalsocietypublishing.org/doi/pdf/10.1098/rsbl.2021.0220>. URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rsbl.2021.0220>.