

# Effet des perturbations humaines sur l'alimentation des narvals

## Sommaire

<b>1 Modélisation de l'effet de l'exposition sur le taux d'émission</b>	<b>2</b>
1.1 Taux d'émission de buzz sans exposition . . . . .	2
1.2 Effet de l'exposition . . . . .	3
1.3 Construction d'intervalles de confiance . . . . .	3
<b>2 Résultats</b>	<b>4</b>
2.1 Recherche de la mémoire optimale . . . . .	4
2.1.1 Note sur le temps d'ajustement des modèles . . . . .	5
2.1.2 Régression bi-exponentielle sur les coefficients autorégressifs . . . . .	5
2.1.3 Coefficients de profondeur . . . . .	6
<b>3 Estimation de l'effet de l'exposition sur le buzzing</b>	<b>6</b>
3.1 Modélisation . . . . .	6
3.2 Intervalles de confiance des coefficients d'exposition . . . . .	8
3.2.1 Variation des coefficients selon des lois normales univariées . . . . .	8
3.2.2 Variation des coefficients selon une loi normale multivariée . . . . .	10
3.3 Comparaison de l'estimation du taux d'émission de buzz avec et sans la profondeur . . . . .	16

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 les activités minières. Les biologistes se questionnent sur les effets potentiels engendrés sur les comportements des narvals.

Afin d'anticiper ces possibles changements de comportements une étude a été conduite pendant plusieurs mois en 2018. Dans ce cadre, 8 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-elle être calculée grâce aux puces GPS placées sur les narvals. Ainsi, nous pouvons tenter de modéliser l'effet de l'exposition des perturbations humaines sur la capacité des narvals à se nourrir.

# 1 Modélisation de l'effet de l'exposition sur le taux d'émission

## 1.1 Taux d'émission de buzz sans exposition

Etant donné que nous voulons modéliser un taux, la régression de Poisson est l'outil tout indiqué. On suppose que le taux d'émission de buzz  $Y$ , dépendant d'un ensemble de variables  $X$ , suit une loi de Poisson de paramètre  $\lambda$ , et on a que  $\mathbb{E}(Y|X) = \lambda$ . La régression de Poisson permet de modéliser le *log* de  $\lambda$  par une combinaison linéaire des variables  $X$  :

$$\log(\lambda) = \beta_0 + X\beta$$

Les variables  $X$  utilisées dans le modèle sont détaillées dans la suite de cette section. Il n'existe pas d'expression explicite de l'estimateur du maximum de vraisemblance des paramètres  $\beta$ , cependant maximiser cette vraisemblance est un problème d'optimisation convexe qui peut être résolu numériquement.

Les données que nous utilisons correspondent à plusieurs individus, nous avons donc plusieurs observations par individus et celles-ci ne sont pas indépendantes. Pour palier ce défaut de modélisation et tenir compte de la spécifité des individus, nous avons ajouté des effets aléatoires  $b_i$  au modèle :

$$\log(\lambda_i) = \beta_0 + b_i + X_i\beta$$

où  $i$  dénote les observations de l'individu  $i$  et  $b_i \sim \mathcal{N}(0, \sigma_i^2)$ ,  $\sigma_i^2$  étant la variance intra-individu.

Pour se nourrir, les narvals doivent plonger profondément (plusieurs centaines de mètres), alors que le reste du temps elles restent "proche" (quelques dizaines de mètres) de la surface. Il faut donc inclure au modèle la 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. Nous pouvons expliciter un peu plus l'expression précédente du modèle :

$$\log(\lambda_i) = \beta_0 + b_i + \text{spline}(\text{Depth}_i)\beta_D$$

Passer par l'estimateur du maximum de vraisemblance nécessite que les observations  $Y_{ij}$  soient indépendantes. Or, 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 rendre les  $Y_{ij}$  indépendants. Pour cela nous introduisons  $K$  variables binaires d'auto-régression codant l'émission d'un buzz aux instants  $t-k$ ,  $k \in \{1, \dots, K\}$ . Le modèle résultant s'écrit alors :

$$\log(\lambda_i(t)) = \beta_0 + b_i + \text{spline}(\text{Depth}_i(t))\beta_D + \sum_{k=1}^K \alpha_k Y_i(t-k)$$

Cette approche nécessite 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 ; 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 une démarche permettant de restreindre le pas de recherche au fur et à mesure que l'on s'approche de  $K_{opt}$  :

1.  $from.k = K_{min}$  ;  $to.k = K_{max}$
2. Tant que  $(to.k - from.k > 2)$  :
  1. Pour  $K_i = from.k + (i - 1) * \lfloor \frac{to.k - from.k}{M-1} \rfloor$ ,  $i \in \{1, \dots, M\}$  :
    1. ajustement d'un modèle avec  $K_i$  éléments mémoire
    2. calcul du BIC
  2.  $i_{opt} = argmin BIC_i$
  3.  $K_{opt} = K_{i_{opt}}$
  4.  $from.k = K_{i_{opt}-1}$  ;  $to.k = K_{i_{opt}+1}$

Le nombre de composants auto-régressifs pouvant être grand, nous avons utilisé une régression bi-exponentielle double permettant de lier le décalage mémoire et le coefficient associé :

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

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.

## 1.2 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.

Afin d'observer l'effet de l'exposition aux perturbations par rapport à l'émission de buzz sans perturbation, nous ajoutons un terme d'"offset". Cet offset est calculé à partir des coefficients  $\hat{\beta}_D$ ,  $\hat{A}_1$ ,  $\hat{lrc}_1$ ,  $\hat{A}_2$ ,  $\hat{lrc}_2$  estimés sans perturbation :

$$offset_i(t) = spline(Depth_i(t))\hat{\beta}_D + \sum_{k=1}^{K_{opt}} (\hat{A}_1 e^{-e^{\hat{lrc}_1} k} + \hat{A}_2 e^{-e^{\hat{lrc}_2} k}) Y_i(t - k)$$

Le modèle incluant l'exposition s'écrit donc :

$$\log(\lambda_i(t)) = offset_i(t) + \beta_0 + b_i + spline(Expo_i(t))\beta_E$$

Et les paramètres estimés  $\hat{\beta}_E$  permettront d'interpréter l'effet de l'exposition par rapport à des conditions de stress normales.

## 1.3 Construction d'intervalles de confiance

Pour se rassurer quant à l'interprétabilité des résultats obtenus il est important de construire les intervalles de confiance des estimations des coefficients associés à l'exposition aux perturbations. 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'auto-régression ne serait

alors pas prise en compte car tuée par l'utilisation de l'offset. Nous avons donc utilisée une approche Monte-Carlo lors de laquelle nous répétons  $M = 1000$  fois l'ajustement du modèle incluant l'exposition en tirant aléatoirement les coefficients  $\hat{\beta}_D$ ,  $\hat{A}_1$ ,  $lrc_1$ ,  $\hat{A}_2$ ,  $lrc_2$  selon leur moyenne et leur variance. Les 1000 ajustements obtenus nous fournissent ainsi un échantillon pour chacun des coefficients d'exposition et nous construisons leur intervalle de confiance via les quantiles empiriques de leur distribution.

## 2 Résultats

### 2.1 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. La figure 1 permet de voir que cet optimal semble bien correspondre à un minimum global.

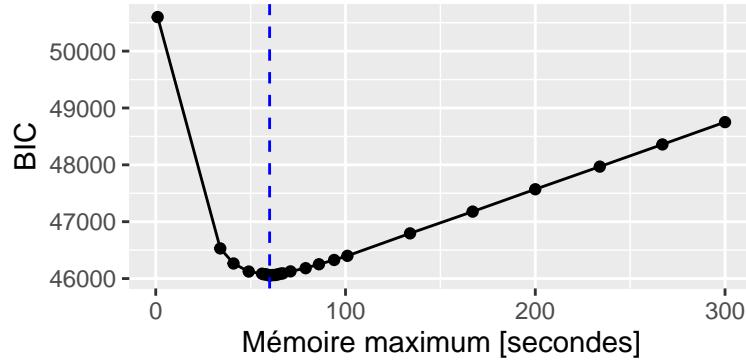


Figure 1: BIC en fonction de la mémoire maximum

Cela nous a permis de restreindre tout de suite l'ensemble de recherche initial à  $\{30, \dots, 90\}$  quand nous avons considéré le modèle incluant les effets aléatoires. De même que précédemment, nous obtenons une mémoire optimale de 60 secondes et la figure 2 nous indique qu'il s'agit bien d'un optimum global.

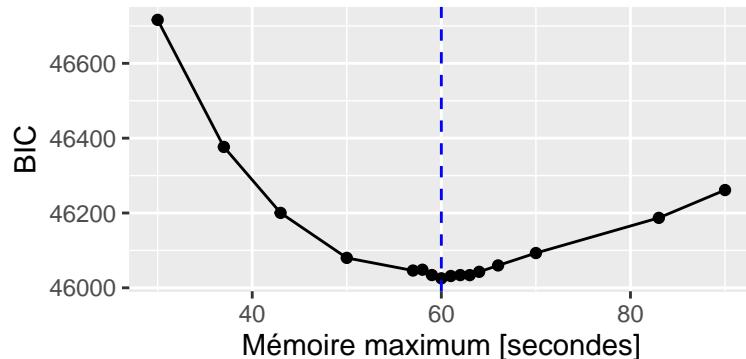


Figure 2: BIC en fonction de la mémoire maximum - modèle mixte

### 2.1.1 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 augmentent.

Comme nous pouvons le voir sur la figure 3 cette augmentation est linéaire pour les modèles classiques, alors que pour les modèles mixtes celle-ci est quadratique.

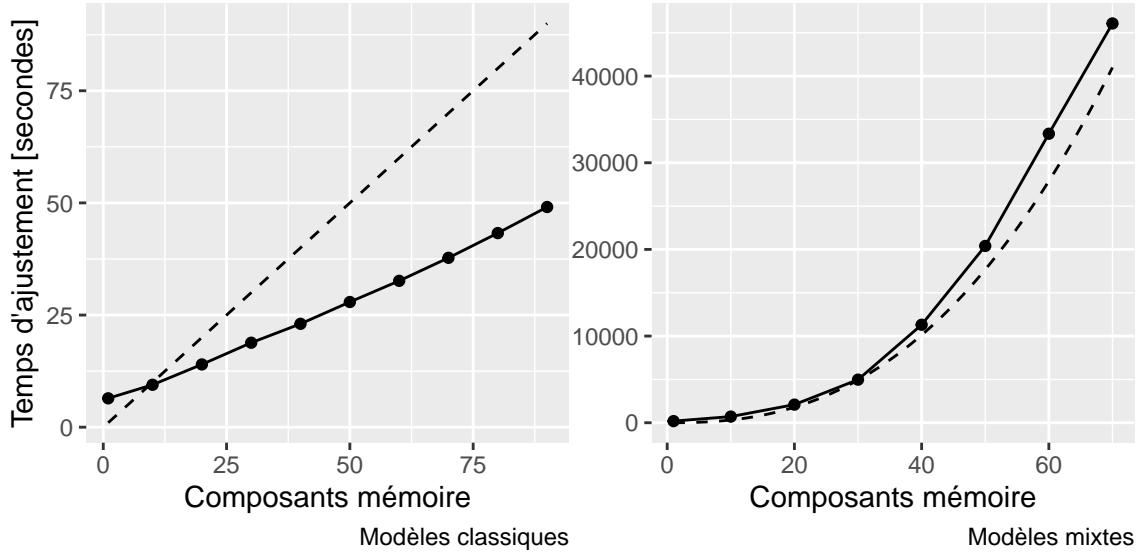


Figure 3: Temps d'ajustement (en secondes) des modèles en fonction de la mémoire maximum

### 2.1.2 Régression bi-exponentielle sur les coefficients autorégressifs

Nous avons vu que les modèles mixtes sont bien plus longs à ajuster que les modèles classiques, aussi, afin de réduire le nombre de variables et le temps d'ajustement, nous appliquons une régression bi-exponentielle aux coefficients auto-régressifs obtenus précédemment. Ainsi nous passons de 60 à 4 variables.

La figure 4 illustre les composantes de la mémoire ajustées pour un décalage maximum de 60. Les points sont les contributions individuelles de chaque décalage, la courbe correspond à la régression bi-exponentielle.

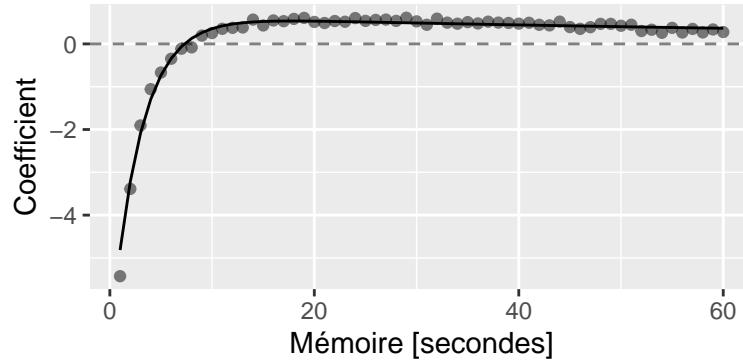


Figure 4: Régression bi-exponentielle des coefficients auto-régressifs

Les coefficients obtenus suite à la régression bi-exponentielle sont présentés dans la table 1.

Table 1: Coefficients autorégressifs obtenus par régression bi-exponentielle

term	estimate	std.error	statistic	p.value
A1	-7.7413301	0.3287556	-23.54737	0
lrc1	-1.0545001	0.0407916	-25.85088	0
A2	0.6544679	0.0427976	15.29218	0
lrc2	-4.6329847	0.1880983	-24.63066	0

### 2.1.3 Coefficients de profondeur

Les coefficients de profondeur du modèle linéaire mixte sans exposition sont proposés table 2.

Table 2: Coefficients de profondeur

term	estimate	std.error	statistic	p.value
splineDepth1	0.0265527	0.1308616	0.2029069	0.8392078
splineDepth2	-0.1254979	0.2069123	-0.6065272	0.5441647
splineDepth3	-23.4762851	2.4222863	-9.6917878	0.0000000
splineDepth4	-32.6394721	3.3383763	-9.7770500	0.0000000

## 3 Estimation de l'effet de l'exposition sur le buzzing

### 3.1 Modélisation

L'exposition aux perturbations est exprimée par  $1/dist$  où  $dist$  est la distance en kilomètres séparant l'animal du bateau dont émane la perturbation.

Nous avons ajusté le modèle mixte suivant :

$$Buzz \sim (1|Ind) + offset(ARDepth) + spline(Expo)$$

Les coefficients estimés par le modèle mixte sans exposition sont fixés dans celui-ci afin que les coefficients ajustés pour la variable d'exposition s'interprètent comme un effet par rapport au comportement normal (sans exposition). De plus, cela permet de réduire sensiblement le coût calculatoire d'ajustement.

Table 3: Coefficients du modèle mixte incluant l'exposition

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	-4.5851696	0.0350968	-130.6436	0
fixed	NA	ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))1	-1.1458123	0.0021544	-531.8501	0
fixed	NA	ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))2	-58.8723530	0.0543014	-1084.1782	0
fixed	NA	ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))3	-112.7904686	0.1069174	-1054.9306	0
ran_pars	Ind	sd__(Intercept)	0.0859693	NA	NA	NA

Table 4: Statistiques du modèle mixte incluant l'exposition

sigma	logLik	AIC	BIC	deviance	df.residual
1	-34393734001	68787468012	68787468075	51675872055	2367453

La figure 5 présente plusieurs graphiques utiles pour valider visuellement le modèle : nous sommes satisfaits de l'absence de corrélation entre les résidus constatée sur les deux premiers et du comportement gaussien visible sur le dernier.

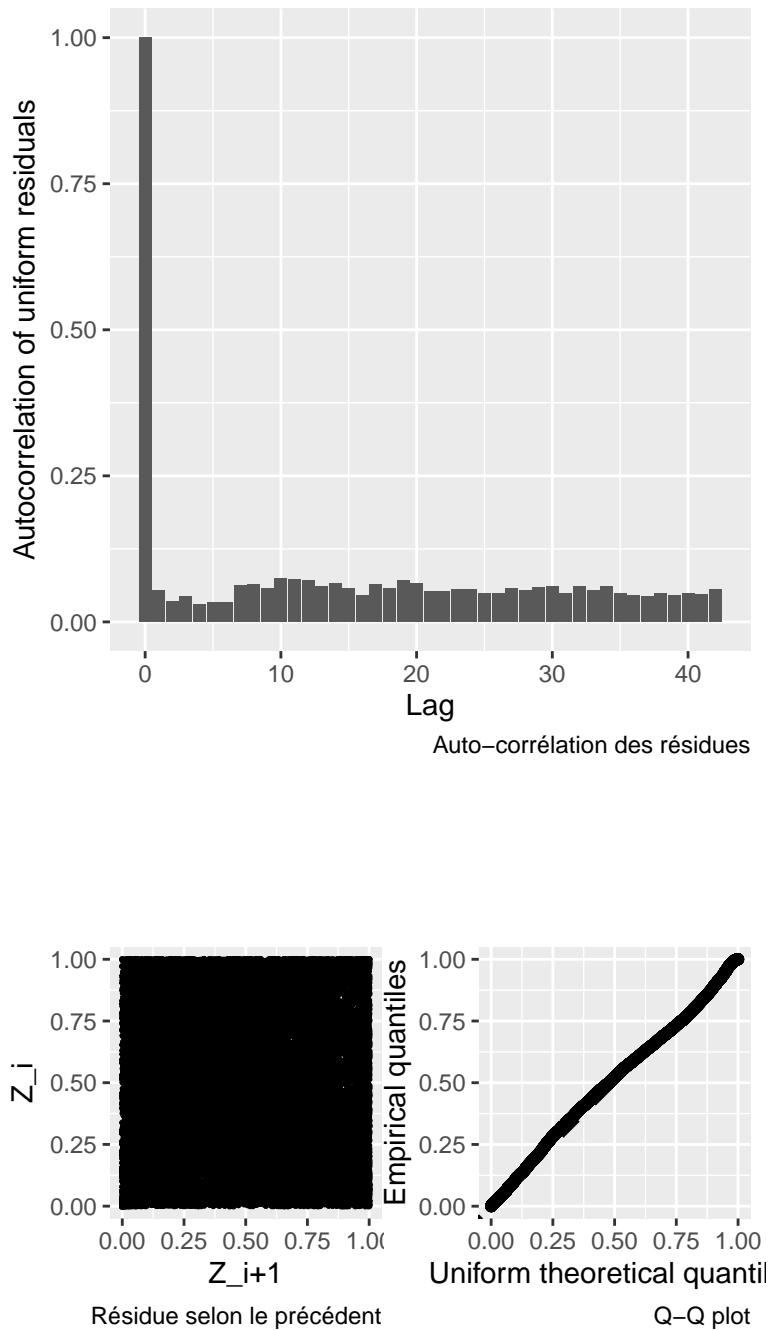


Figure 5: Validation graphique du modèle

## 3.2 Intervalles de confiance des coefficients d'exposition

Etant donné que nous avons fixé les coefficients associés à la profondeur et à l'effet mémoire aux valeurs obtenues sans exposition, l'estimation des intervalles de confiance des coefficients d'exposition ne peuvent être construits directement à partir de leurs erreurs standards, car la variabilité des coefficients de profondeur et d'effet mémoire ne serait pas prise en compte. Nous avons donc mis en place une approche Monte Carlo ajustant 1000 fois le modèle avec exposition tout en faisant varier les coefficients fixés, et permettant finalement d'estimer les intervalles de confiance des coefficients d'exposition via les quantiles des distributions obtenues.

### 3.2.1 Variation des coefficients selon des lois normales univariées

Dans un premier temps, nous avons considéré que les coefficients 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 biexponentielle et 4 autres pour l'interpolation polynomiale par morceaux sur la profondeur.

Sur la figure 6 nous avons représenté les courbes des fonctions biexponentielles ainsi générées. Nous pouvons voir que leurs allures semblent toujours suivre celle de la régression initiale.

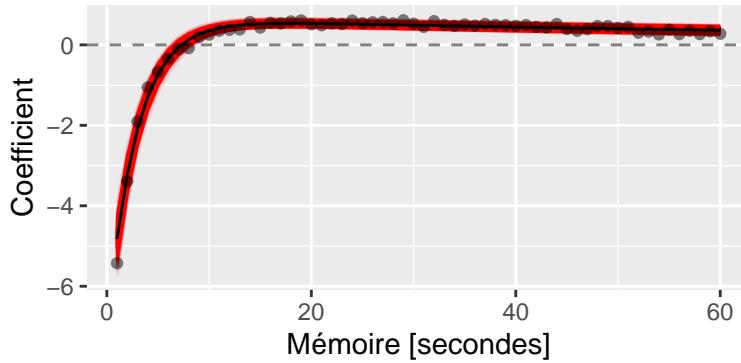


Figure 6: Variation des coefficients autorégressif selon des lois normales univariées

Nous avons fait de même avec les coefficients de profondeur, et nous pouvons remarquer sur la figure 7 que cette fois-ci l'allure de certaines courbes s'éloignent sensiblement de l'interpolation polynomiale moyenne.

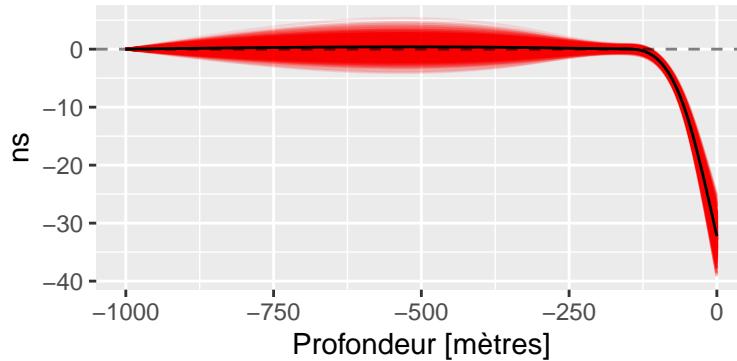


Figure 7: Variation des coefficients de profondeur selon des lois normales univariées

Les intervalles de confiance calculés avec la procédure Monte Carlo sont donnés sur la table 5 (colonnes préfixées par “MC”). 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 (colonne préfixées par “SE”). 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 l’émission de buzz.

Table 5: Intervalles de confiance dans le cas de normales univariées

	MC - inf	MC - sup	SE - inf	SE - sup
(Intercept)	-6.441511	-2.968321	-4.638673	-4.494102
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))1	-11.898716	11.294079	-1.430436	-1.422075
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))2	-152.959472	1.047762	-56.014702	-55.803478
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))3	-306.428946	11.706679	-106.846662	-106.430830

La figure 8 donne une représentation graphique de ces intervalles (en rouge les “MC” et en bleu les “SE”), ainsi que des distributions des coefficients.

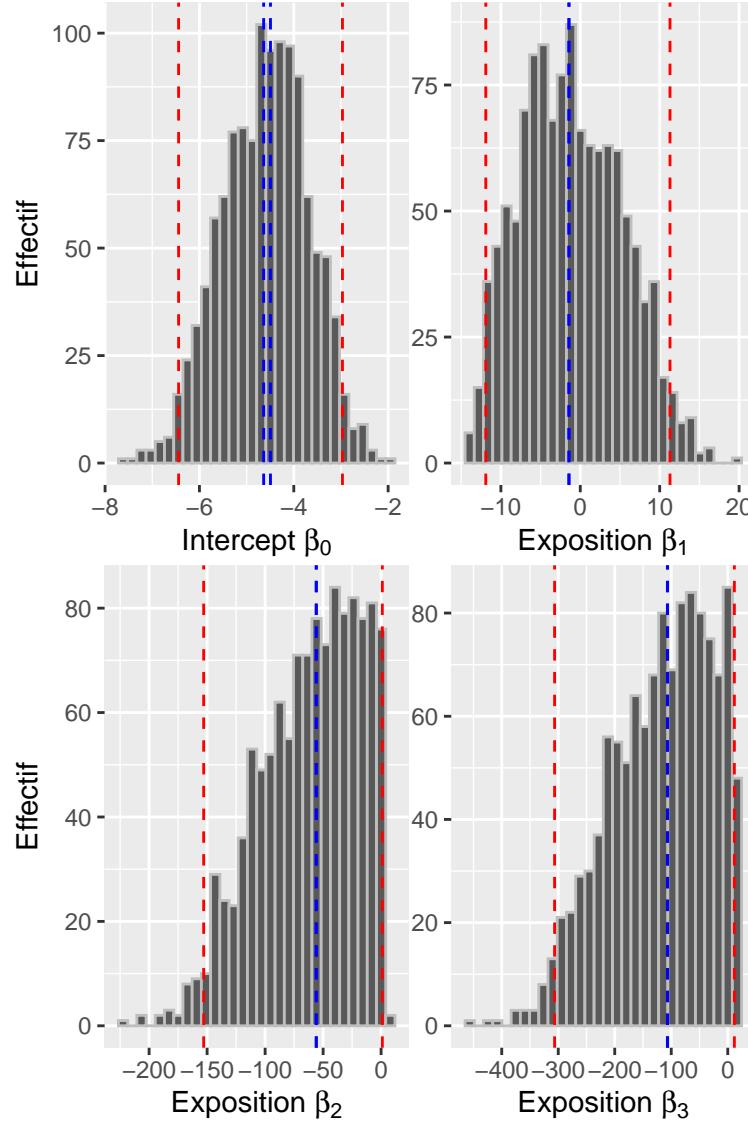


Figure 8: Intervalles de confiance dans le cas de normales univariées

### 3.2.2 Variation des coefficients selon une loi normale multivariée

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 des lois normales multivariées (une pour les coefficients de la régression biexponentielle et une pour les coefficients de l'interpolation polynomiale) toujours centrées sur les estimations moyennes, mais ayant des matrices de variance-covariance non diagonales. Dorénavant nous avons donc une loi normale multivariée pour les coefficients de la régression biexponentielle et une autre pour la profondeur, les 2 étant de dimension 4.

La figure 9 montre que la régression biexponentielle initiale est encore plus fidèlement suivie qu'auparavant.

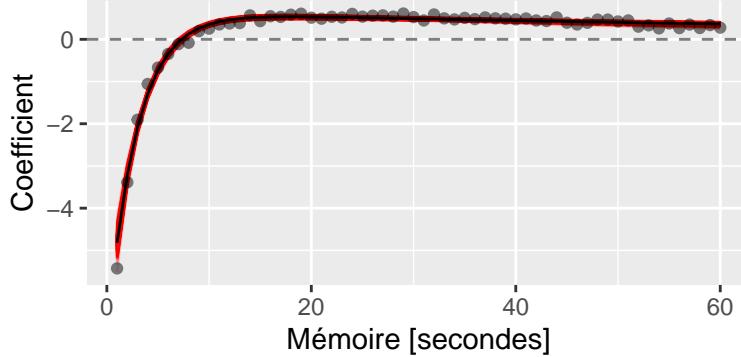


Figure 9: Variation des coefficients autorégressifs selon une loi normale multivariée

Et surtout, comme nous pouvons le voir sur la figure 10, il en va de même pour la profondeur, alors que précédemment les tirages donnaient des courbes fortement éloignées de celle attendue.

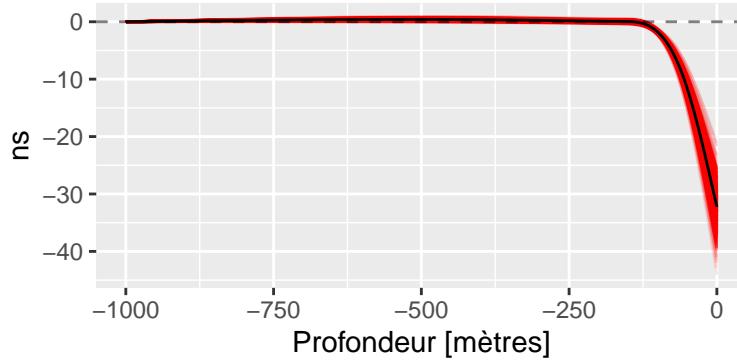


Figure 10: Variation des coefficients de profondeur selon une loi normale multivariée

Sur la table 6 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 l'émission de buzz.

Table 6: Intervalles de confiance dans le cas de normales multivariées

	MC - inf	MC - sup	SE - inf	SE - sup
(Intercept)	-4.887771	-4.273240	-4.631902	-4.497720
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))1	-1.971359	-0.480328	-1.237372	-1.228926
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))2	-64.614962	-52.023774	-58.309579	-58.096519
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))3	-124.492842	-98.741635	-111.627518	-111.208774

Nous pouvons également voir sur la figure 11 que les distributions semblent normales, contrairement à celles observées avec les lois univariées.

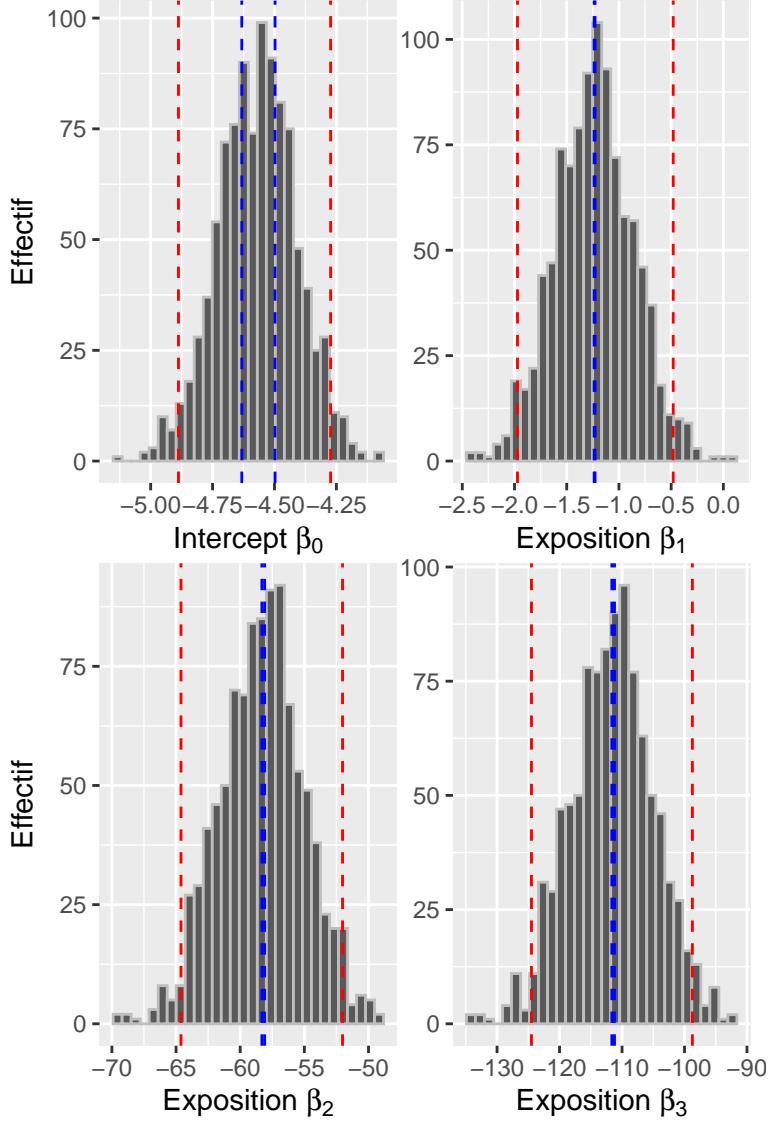


Figure 11: Intervalles de confiance dans le cas de normales multivariées

### 3.2.2.1 Variance-covariance des coefficients autorégressifs et non des coefficients de la régression biexponentielle

Bien que plus intéressante, la procédure que nous avons mise en place se base sur la matrice de variance-covariance des coefficients de la régression biexponentielle, ce qui signifie que nous ne captions pas directement la variabilité du phénomène autorégressif mais plutôt celle de la régression biexponentielle. Pour palier cette approximation nous avons tenté d'estimer la matrice de variance-covariance des coefficients autorégressifs, à nouveau par une procédure Monte Carlo. Cette fois-ci nous répétons 1000 fois le tirage des 60 coefficients de mémoire à partir d'une loi normale multivariée et à chaque fois nous ajustons une régression biexponentielle sur les 60 coefficients tirés ; finalement nous pouvons obtenir les estimations moyennes des coefficients de la biexponentielle, ainsi que la matrice de variance-covariance associée. Nous reprenons enfin la procédure Monte Carlo employée dans la section précédente avec le vecteur de moyennes et la matrice de variance-covariance obtenus.

Nous pouvons contrôler sur la figure 12 que les régressions biexponentielles obtenues sont cohérentes.

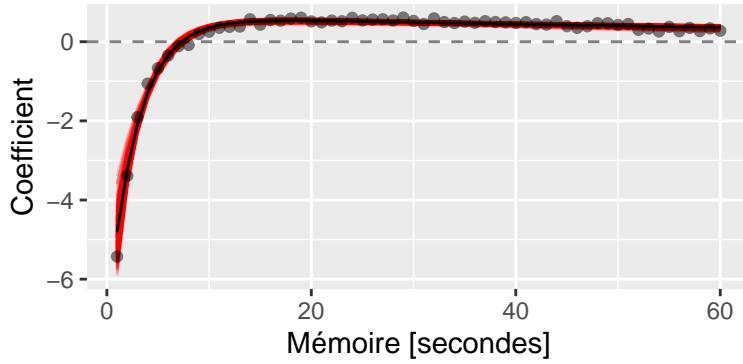


Figure 12: Variation des coefficients autorégressif selon une loi normale multivariée estimée par Monte Carlo

La table 7 et la figure 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 7: Intervalles de confiance dans le cas de normales multivariées

	MC - inf	MC - sup	SE - inf	SE - sup
(Intercept)	-4.865849	-4.2754368	-4.645331	-4.510474
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))1	-1.981092	-0.4261478	-1.122103	-1.113660
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))2	-65.140058	-51.9494486	-59.342551	-59.127834
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))3	-125.571605	-98.5452180	-113.695413	-113.273788

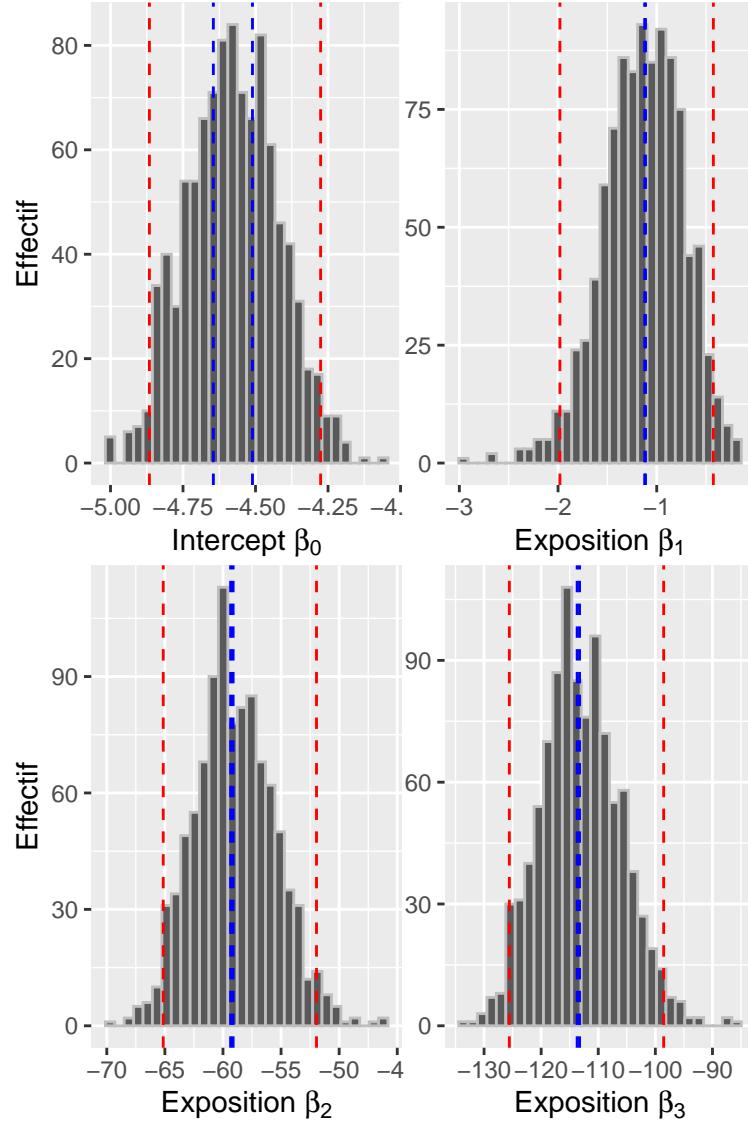


Figure 13: Intervalles de confiance dans le cas de normales multivariées

### 3.2.2.2 Sans passer par la régression biexponentielle

L'intérêt de la régression biexponentielle était 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. Ainsi, nous avons tenté de nous passer de la régression biexponentielle, et donc de tirer les 60 coefficients mémoire et les 4 coefficients de profondeur directement dans une seule loi normale multivariée.

Nous pouvons voir sur la table 8 et la figure 14 qu'avec cette approche plus directe, les intervalles de confiance sont quasiment identiques à ceux obtenus dans la section 3.2.2, et même plus petit pour l'intercept.

Table 8: Intervalles de confiance dans le cas de normales multivariées

	MC - inf	MC - sup	SE - inf	SE - sup
(Intercept)	-4.877449	-4.2607168	-4.654712	-4.512880
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))1	-1.782223	-0.5655889	-1.186667	-1.178238
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))2	-64.410381	-52.0926295	-58.180603	-57.969085
ns(X, knots = quantile(data\$X[data\$X > 0], 1:2/3))3	-124.003446	-99.0561959	-111.400643	-110.983884

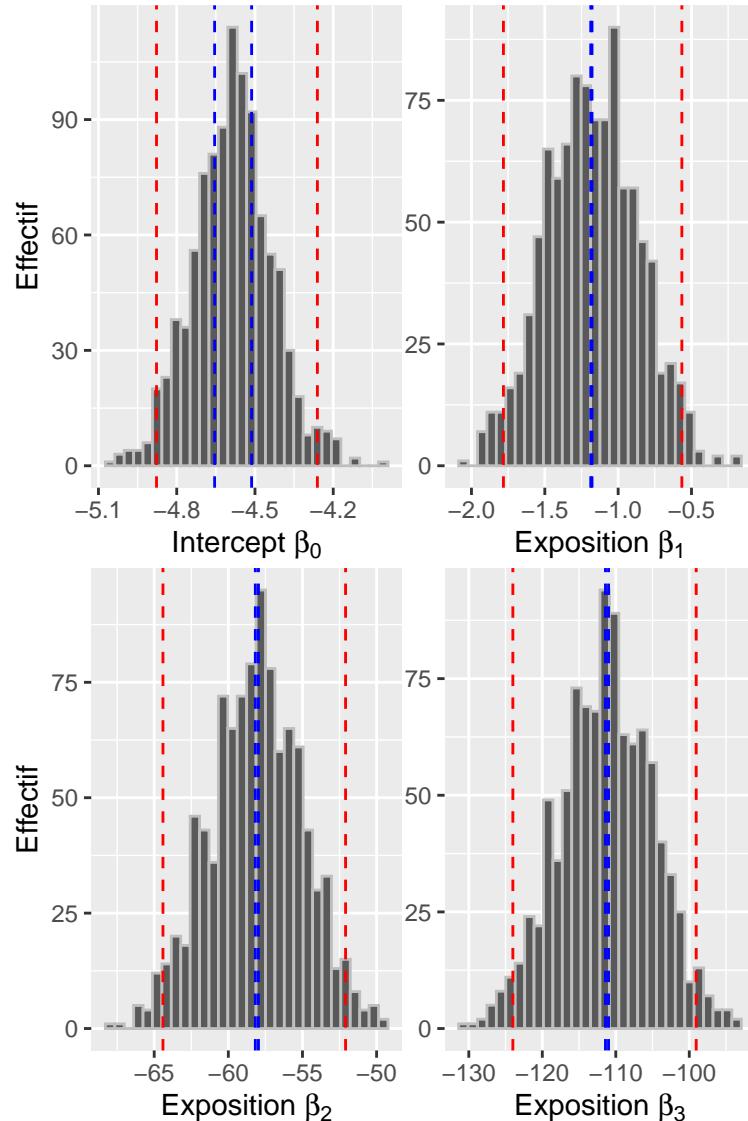


Figure 14: Intervalles de confiance dans le cas de normales multivariées

### 3.3 Comparaison de l'estimation du taux d'émission de buzz avec et sans la profondeur

