Analysis of Greenland narwhals's behavioural responses to ship noise and airguns exposure using varying coefficients correlated velocity models

Alexandre Delporte* [1], Susanne Ditlevsen†[2], and Adeline Samson‡[3]   + Mads Peter ?

[1] *Laboratoire Jean Kuntzmann, Université Grenoble-Alpes, France*
[2] *Department of Statistics, University of Copenhagen, Denmark*
[3] *Laboratoire Jean Kuntzmann, Université Grenoble-Alpes, France*

Dated: April 15, 2024

Tittle Varying coefficients correlated velocity models
- with complex boundary constraints
- with ~~boundary~~ constraints due to complex landscapes
- in complex landscapes

with application to narwhal behavioral response to ship noise and airgun exposure

Journal : Methods in ecology ?

Data : - GPS location or interpolation ?
- don't remove points on land, they should be modeled properly thanks to the measurement noise ?

Figure : il y en a trop pour le moment. Il faut réussir à réduire à 4 ou 5 maximum

*alexandre.delporte@univ-grenoble-alpes.fr
†susanne@math.ku.dk
‡adeline.leclercq-samson@univ-grenoble-alpes.fr

En anglais, pas d'espace avant une ponctuation (:, ! ?)

1 Introduction

Anthropogenic underwater noise in the Arctic environment causes disturbances in the endemic wildlife. More precisely, marine mammals, whose vital functions highly depend on sound perception for communication and orientation, are likely to react to such disturbances. The Arctic Council report [**Halliday, Pine, and Insley 2020**] emphasized the need for more knowledge about the impact of underwater noise on Arctic marine mammals. The main sources of noise in the region are vessel traffic, ice breakers, drilling, sonars and seismic airguns used for oil and gas exploration. Among those sources, an increase in vessel traffic is anticipated due to the melting of the ice cap, and seismic airguns are reported to have the largest impact on underwater noise. Two types of consequences can be considered : injury and behavior disturbance. While physical harm is straightforward to assess and mitigate by determining species-specific received sound level thresholds, no canonical method is expected to be found to examine behavior disturbances, due to the multiplicity of contextual variables that can influence a behavioral reaction, and the variety of these reactions.
Research recommendations include a particular focus on measurement of behavioral reactions to various type of sound exposure [**Southall, Bowles, et al. 2008**]; [**Southall, Finneran, et al. 2019**]. Rich behavioral studies typically involve controlled exposure experiments during which animals are exposed to disturbances in a precise and monitored set up where different statistics can be measured, such as GPS positions, depth, sound exposure level, temperature, pressure, heart beat, stroke rate, vocalizations.
Numerous surveys have been conducted about beaked whales, which have been regularly implicated in stranding events following sonar exercises [**Tyack et al. 2011**]; [**Cioffi et al. 2022**], belugas [**Martin et al. 2023**], sperm whales [**Madsen et al. 2006**], blue whales [**Friedlaender et al. 2016**], or narwhals [**Heide-Jørgensen et al. 2021**]; [**Tervo et al. 2021**] ~~and many other marine mammals.~~
A severity scale was defined in [**Southall, Bowles, et al. 2008**] to help assessing which behavioral responses are more likely to alter a population's capacity to survive, reproduce or forage. Changes in movement speed and direction, avoidance reactions as well as modified dive profiles or vocalizations were ranked between 4 and 7 on a scale out of 9 (9 being stranding events that can directly lead to the death of the animal) depending on the magnitude of the changes and their duration.

So far, the studies about narwhals behavioral responses to sound exposure proved that they exhibit avoidance reactions and have a proclivity to move toward the shore when exposed to seismic airguns [**Heide-Jørgensen et al. 2021**]. Moreover, a significant decrease in their buzzing rate has been assessed as far as 40 km away from the sound source [**Tervo et al. 2021**]. These results are based on controlled exposure experiments conducted in 2017 and 2018 in South-east Greeland, which is home to a declining population of narwhals [**Garde et al. 2022**]. Resulting GPS positions of the narwhals have been analyzed with discrete time methods, typically relying on Markov chains. Since GPS positions are retrieved irregularly in time, each narwhal track was linearly interpolated to obtain a position every second, and the positions were then classified in three states "Far from the shore", "Moving toward shore" and "Close to the shore" at each second. An exposure covariate defined as the inverse of the distance to the sound source was used to model the transition probabilities of the Markov Chain, and it was found that the probabilities of being in states "Close to the shore" and "Moving toward the shore" increased with exposure to the sound, while probability of being in state "Far from the shore" decreased. In this paper, we adopt a different approach to describe horizontal movement and make use of a continuous time model to analyse the GPS positions. The objective is to be able to quantify the effect of sound exposure on different characteristics of the movement.

The dynamic of the states was model with a Markov chain, that assuming the current state only depends on the last one in time.

aller à la ligne.

We focus on continuous time models defined as the solution of a SDE.

A stochastic differential equation with time-varying coefficient is employed to describe the GPS tracks of the individuals [**Michelot, Glennie, Harris, et al. 2021**]. One widely known model for such data is the continuous-time correlated random walk, also known as the integrated Ornstein-Uhlenbeck or correlated velocity model (CVM) [**Johnson et al. 2008**]. It has been used extensively to analyze animal

quelle est la logique de l'ordre des refs ici?

- On ne commence pas une phrase par une notation ou un symbole mathématique.
- Parfois tu utilises du passé, du présent ou du futur - Ce n'est pas homogène -

movement in two or even three dimensions [**Johnson et al. 2008**]; [**Gurarie, Fleming, et al. 2017**]; [**Alt and Hoffmann 1990**]; [**Gurarie, Grünbaum, and Nishizaki 2011**]; [**Albertsen 2018**], ~~and~~ This model can be viewed as a special case of velocity potential model, where the potential is a second degree polynomial [**Preisler, Ager, and Wisdom 2013**]. The following formulation of the equations is based on biologically interpretable parameters [**Gurarie, Fleming, et al. 2017**]:

$$\begin{cases} dX(t) = V(t)dt \\ dV(t) = -\begin{pmatrix} \frac{1}{\tau} & -\omega \\ \omega & \frac{1}{\tau} \end{pmatrix} (V(t) - \mu)dt + \frac{2\nu}{\sqrt{\pi\tau}} dW(t) \end{cases} \quad (1)$$

where $V(t)$ is the horizontal ~~2D~~ dimensional velocity at time t and $X(t)$ is the position, typically in longitude and latitude or in UTM coordinates, $W(t)$ is a 2-dimensional brownian motion. The parameter τ is an autocorrelation time scale, ν controls the norm of the velocity and drives ~~how much~~ the random variability ~~there is in the velocity vector~~, μ is the long term velocity, while ω is an angular velocity that controls how fast the velocity vector rotates. Whether or not to include a non-zero mean velocity parameter μ depends on the specific context of the study. Typically, it would make sense to have $\mu \neq (0,0)$ when examining migratory patterns or avoidance reactions.

Constraints for the shore are needed in the model.

In general, the motion of marine mammals is subject to spatial constraints induced by the coastline. This is particularly the case for the narwhal data we want to analyse: the animals move in a complex system of fjords in Scoresby Sound (East Greenland), imposing strong constraints on the location process. However, this is not the case ~~this phenomenon is not taken into account in model~~ 1 ~~nor~~ in most of the SDE used ~~in~~ for ~~the context of~~ animal movement [**Johnson et al. 2008**]; [**Michelot, Glennie, Harris, et al. 2021**]; [**Gurarie, Fleming, et al. 2017**]. ~~Recent research has been conducted to include constraints in these type of models.~~ Recently, A reflected version of the CVM has been considered to study sea lion telemetry data [**Hanks, Johnson, and Hooten 2017**]. Simulation of constrained paths was achieved by projecting on the coastline the points that eventually reached land. Estimation of the SDE parameters relied on computationally intensive MCMC algorithm (Metroplis-Hastings), and required high frequency data. ~~This requirement was met by linearly interpolating between the discrete-time observations of the process.~~ Another idea is to add a repulsive potential function in the drift, ~~This has been~~ as done to constrain ants movement within a box [**Russell et al. 2018**], ~~but requires very~~ Only simple boundaries are possible in order to get an explicit expression of the potential function. This is not the case with fjord systems.
Moreover, the way to model the motion of an animal near the boundary of the domain should be species-specific, accounting for the particular behaviors and spatial use of the studied animal [**Brillinger 2003**]. Thus there is a need to develop new models adapted to complex constraints
For narwhals, it was observed from the data that they have a proclivity to move along the shoreline. Here we propose a new model based on the following ~~simple~~ idea: as the narwhals approach the shore, at some point they need to rotate to avoid or align with the shoreline. The novelty of this paper consists in modeling ~~estimating~~ the parameter ω as a smooth function of the distance to the boundary and the angle between the animal's direction and the boundary normal vector, ~~therefore~~ this makes the drift in the velocity equation dependent on the location process x and the boundary of the domain. Note that in this framework, the special case $\omega = 0$ would be the unconstrained correlated velocity model of [**Johnson et al. 2008**].
We show that the angular velocity can be estimated as a function of the two covariates in the framework of the R package `smoothSDE`. Def. To make this possible, we formulated a state-space model by binding the velocity and location in one single hidden state vector, following [**Johnson et al. 2008**]. Maximum likelihood estimation based on Kalman filter is then performed as for any other model in `smoothSDE`. + covariates that depend on the ...

Sound exposure ~~will be~~ is considered to affect only the parameters ν and τ in equation 1. This effect will be estimated as a smooth function of a sound exposure covariate defined as the inverse of the distance to the ship. This choice of covariate is motivated by ~~previous work in~~ [**Heide-Jørgensen et al. 2021**]. A response model will be compared to a baseline model representing the narwhals movement before exposure ?

to be defined

to the disturbance, that is under normal conditions [**Michelot, Glennie, Thomas, et al. 2022**]. The effect of sound exposure is assessed both in the constrained and unconstrained ($\omega = 0$) frameworks.

To summarize, the main contributions of this paper are

à dire avant ?

c'est pas clair ce qui est la contribution de ce papier

- Derivation of a state-space model for the so called "Rotational Advective Correlated Velocity Model" (RACVM, 1) [**Gurarie, Fleming, et al. 2017**], and addition of this model in the framework of smoothSDE R package, to enable the use of smooth parameters depending on external covariates, and estimation from noisy observations irregularly spaced in time.

pas de A ?

- Definition of a Constrained Rotational Correlated Velocity Model (CRCVM) for simulation study. Deviations angles from shoreline and distance to shore are used as covariates to constrain the movement within a polygon, and align the velocity with the boundary of the domain.
- Application to real narwhal data to asess a behavioural response of the narwhals, using the exposure covariate defined as the inverse of the distance to the sound source.

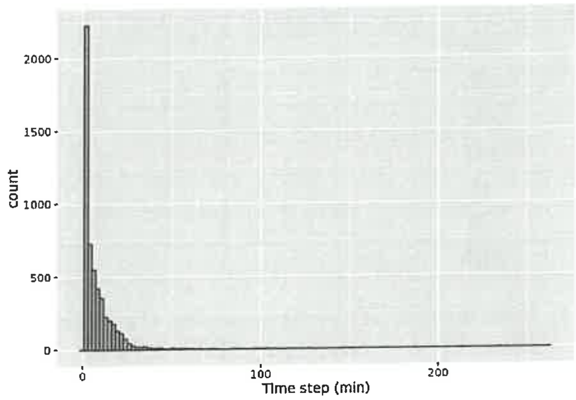
Section 2 gives an overview of the narwhal data available for the analysis. Then in Section 3, we discuss the diffusion models and the introduction of covariates in the diffusion parameters through mixed effect models. Section 4 is dedicated to a comprehensive simulation study, whereas in the last section, the results of the application to the narwhal data are examined.

2 Movement data of Greenland endemic narwhals

2.1 Description of the movement data

~~As already mentioned,~~ The dataset analysed in this paper has been the subject of several studies, focusing mainly on vocalizations and avoidance reactions [**Heide-Jørgensen et al. 2021**]; [**Tervo et al. 2021**]. Six male narwhals were equipped with FastLoc GPS receivers in August 2018 in Scoresby Sound fjord system in South-East Greenland by biologists from the Greenland Institute of natural ressources, with the help of local Inuit hunters. The Scoresby Sound fjord system is known to be the summer residence for an isolated population of narwhals. Here we recall briefly how the data has been retrieved. For more details about the study area and the tagging of the animals, the reader may refer to [**Heide-Jørgensen et al. 2021**]. An offshore patrol vessel military ship was sailed to shoot airguns underwater between August 25 and September 1. It was equipped with two Sercel G-guns at 6m depth and moved at a speed of 4.5 knots. The guns were fired synchronously every 80 s, and the GPS navigation system recorded the location of every shot. Data includes narwhals's latitude and longitude, distance relative to the ship in metres, GPS time, distance to the shore in metres. GPS positions are known only at specific times when the narwhals get close to the sea surface. The median time step between two GPS measurements in the data is about 5 minutes and only 0.3% of the time steps reach more than two hours, with a maximum at more than 4 hours. Figure 1 shows the distribution of time steps in the data.

est ce que tu utilises les données interpolées ?



Mettre fig 1, 2 en 1 seule figure

Figure 1: Histogram of time step between consecutive GPS positions

Experiments were divided into unexposed periods - for which the narwhals are not in line of sight with the ship - trial periods - when the narwhals are exposed to the ship and airguns are shot - and intertrial periods - when the narwhals are exposed to the ship but airguns are not shot. These periods are indicated by a categorical variable T_{ship} in the dataset. Figure 2 shows how these periods are distributed among the 6 narwhals that were tracked.

est-ce que tu utilises ces 2 covariables dans l'analyse? Si tu n'utilises que Z = exposition binaire, il faut modifier cette partie

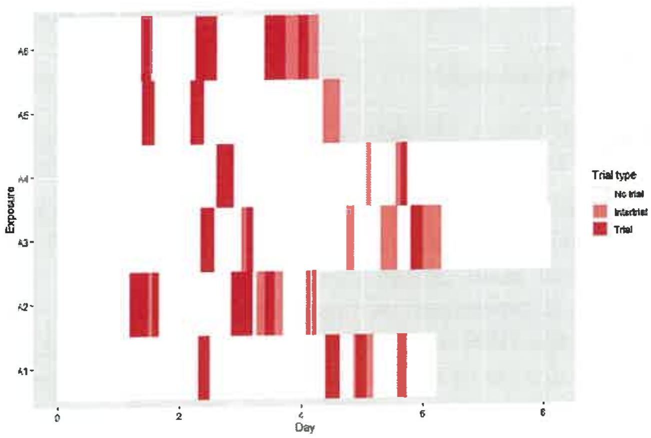


Figure 2: Trial and Intertrial periods for each narwhal

For each narwhal, the entire track was separated into a period before exposure defined as the period before the narwhal gets in line of sight with the ship for the first time, and a period during exposure, after the narwhal has been in line of sight with the ship for the first time. We kept only GPS measures 24 hours after tagging, to avoid any tagging effects on the narwhals's movement, as recommended in [**Heide-Jørgensen et al.** 2021]. Measures that resulted in a velocity higher than 20 km/h were also discarded (only 2 data points for the same narwhal track). Overall, 4192 GPS positions were kept for the analysis and projected to UTM coordinates zone 26. The splitting between data before and after exposure resulted in 935 measures before exposure and 3257 measures after exposure. Table 1 shows how the data is distributed among the different narwhals. Figure 3 shows all the tracks before and during exposure.

en annexe ?

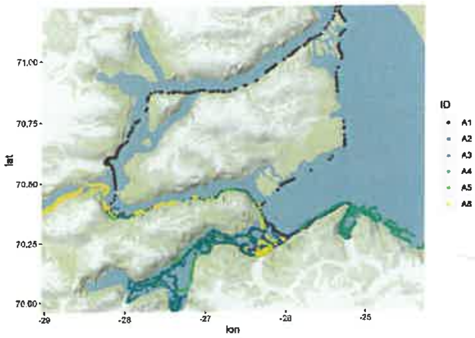
peut-être en appendix ?

Narwhal ID	Number of measurement before exposure	Number of measurements during exposure
A1	258	576
A2	40	515
A3	222	680
A4	291	642
A5	48	419
A6	76	425
Total	**935**	**3257**

Table 1: Distribution of the data among the 6 individuals

pas utile ? en annexe ? ou à enlever ?

(a) Tracks before exposure experiments



(b) Tracks during exposure experiments

Figure 3: Movement data of East Greenland narwhals

2.2 Introduction of new variables describing the constraints in the fjord systems.

The land geometry for this specific region in South-East Greenland was gathered from OpenStreetMap database. Supposing there are n observations, for $j \in \{1, \cdots, n\}$, the GPS position obtained at time t_j is denoted $y(t_j)$, and the nearest point on the shoreline is denoted $p(t_j)$. The distance to the shore, denoted $D_{shore}(t_j)$ in the sequel, is determined as the distance to this point. The resulting values range from 0 to 7.5 km. About 9% of the GPS measurements in the dataset turned out to be on land. This was primarily attributed to inaccuracies in the shoreline maps rather than errors in GPS measurements. For constrained movement models, these points were ignored. Then, the exposure to shore covariate was defined with the following formula

$$\begin{cases} E_{shore}(t_j) = 0 & \text{if } D_{shore}(t_j) > D_0 \\ E_{shore}(t_j) = \frac{1}{D_{shore}(t_j)} & \text{if } D_{shore}(t_j) \leq D_0 \end{cases} \quad (2)$$

where the threshold D_0 was chosen equal to 0.5 km. The angle between the narwhal's heading and the shoreline is estimated as the angle $\Theta(t_j)$ between the observed empirical velocity $\hat{v}(t_j) = \frac{y(t_j+\Delta_j)-y(t_j)}{\Delta_j}$, where $\Delta_j = t_{j+1} - t_j$ and the vector $n(t_j) = y(t_j) - p(t_j)$. This is illustrated in figure 4. A value $\Theta(t_j) = \pm\frac{\pi}{2}$ indicates movement parallel to the shore, while $\Theta(t_j) \in]\frac{\pi}{2}, \pi] \cup [-\pi, -\frac{\pi}{2}[$ indicates movement toward the shore. The distance from the ship, denoted $D_{ship}(t_j)$, was computed from the GPS ship and narwhals's positions. The values are comprised between 2.68 and 63.8 km. Histograms of the distance to the ship and the distance to the shore are shown in figure 5. ~~The covariates E_{shore} and Θ will only be used for the smooth angular velocity parameter ω in the constrained SDE models.~~

à justifier - Plus tard n'utilises 0 3km

$D_{shore} = 0$ ou un seuil petit ?

1 à la ligne

Explique avec les notations y_{ij}, $i=1,\cdots,I$, $j=1,\cdots$ mesure bruitée d'une position - Puis explique que pour le reste du papier, tu as besoin de définir 3 nouvelles variables, Θ et E_{shore}, E_{ship} et ensuite tu les définis

pourquoi à cet endroit

→ et pourquoi ne pas considérer que ce sont des erreurs de GPS et les laisser en mer ...

trop tôt, tu n'en as pas encore parlé

notations à introduire avant
La section 2.2 pourrait contenir :
- notations i, j y_{ij}, Δ, t_{ij},
- y_{ij} comme mesure bruitée de GPS, d'où 9% sont sur la terre
- introduction de $p(t_{ij})$ le pt le + proche (cet définition p de y est sur la mer)
- introduction de Dshore
- Θ
- Dship

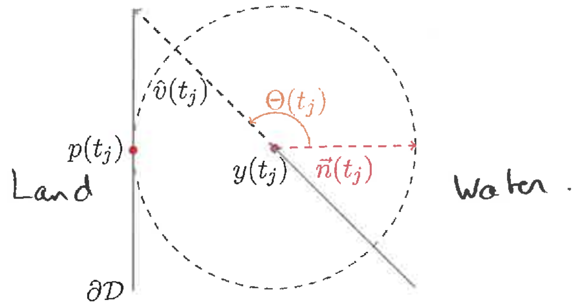
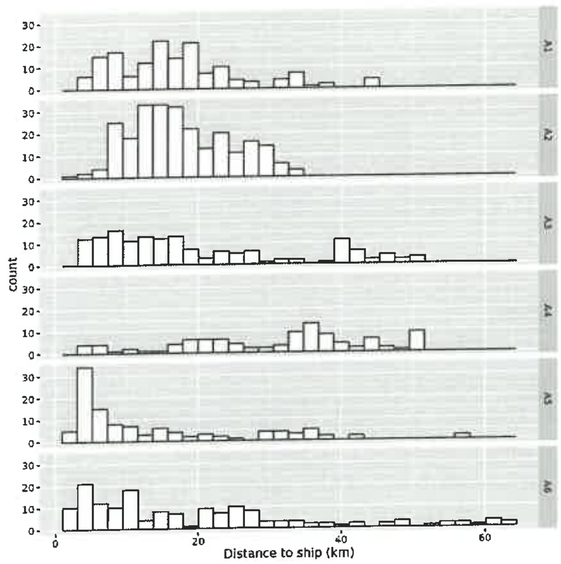
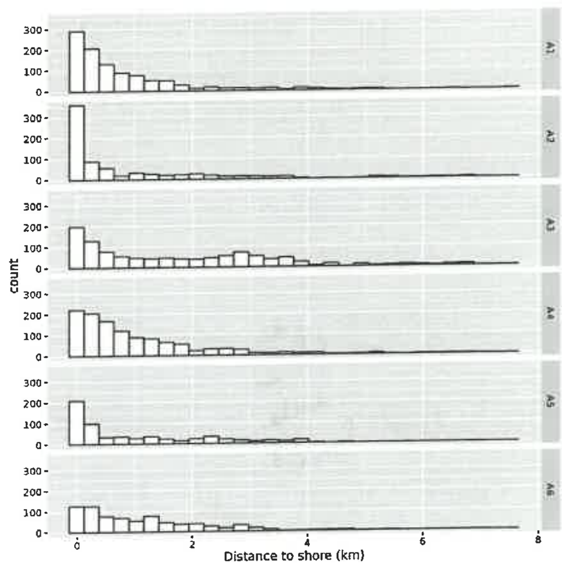


Figure 4: Example of nearest shore point and angle Θ. ∂D represents the boundary of the domain.



(a) Histograms of the distance to the ship for each narwhal



(b) Histograms of the distance to the shore for each narwhal

pas utile ?

plutot fig 6 et son equivalent pour Eshore ?

Figure 5: Distance to ship and distance to shore covariates

dans le même paragraphe que Dship

The exposure to ship covariate, denoted E_{ship}, was defined as the inverse of the distance to the ship (in km) [**Heide-Jørgensen et al.** 2021]. D_{ship} is only ~~known~~ defined when the narwhal is in line of sight with the ~~boat~~ ship, so that exposures are set to 0 when this is not this case. This implies that the effect of the ship noise on the narwhals's movement is only considered in this specific case in the whole study, though it is very likely that narwhals can perceive the disturbance even when they are not clearly in line of sight with the sound source. This covariate is meant to be a proxy for sound exposure levels received by the narwhals. Reasonably, the closer the narwhal is to the ship, the more it can hear its sound, and the greater the exposure. Exposure levels for each narwhals are displayed in figure 6. All the relevant covariates that will be used for the analysis of narwhals movement are summarized in table 2.

à remonter en 2.1 quand tu définis l'exposition ?

This provides a conservative estimate of the effect of the noise exposure

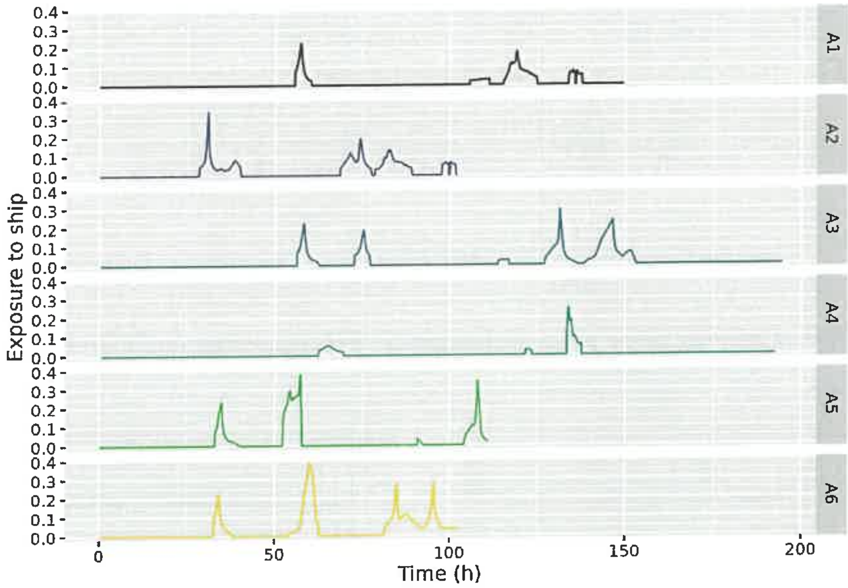


Figure 6

Covariate	Unit	Description	Domain
$D_{ship}(t)$	km	distance in kilometers between the narwhal and the ship at time t	$\mathbb{R}^+$
$T_{ship}(t)$	categorical	0 when there is no trial, 1 if it is an intertrial period, and 2 when it is a trial period	$\{0,1\}2$?
$E_{ship}(t) = \frac{1}{D_{ship}(t)}$ $\mathbb{1}_{T_{ship}\neq 0}$	km^{-1}	global exposure level of the narwhal to the ship disturbance at time t	$\mathbb{R}^+$
$E_{ship,I}(t) = \frac{1}{D_{ship}(t)}\mathbb{1}_{T_{ship}(t)=1}$	km^{-1}	intertrial exposure level of the narwhal to the ship disturbance at time t	$\mathbb{R}^+$
$E_{ship,T}(t) = \frac{1}{D_{ship}(t)}\mathbb{1}_{T_{ship}(t)=2}$	km^{-1}	trial exposure level of the narwhal to the ship disturbance at time t.	$\mathbb{R}^+$
$D_{shore}(t)$	km	distance between the narwhal and the nearest point on the shore at time t	$\mathbb{R}^+$
$E_{shore}(t) = \frac{1}{D_{shore}(t)}\mathbb{1}_{D_{shore}(t)<D_0}$	km^{-1}	global exposure level of the narwhal to the shore at time t	$\mathbb{R}^+$
$\Theta(t)$	rad	angle between the vector that goes from the nearest shore point to the narwhal's position and the empirical step vector at time t	$[-\pi,\pi]$

tu les utilises ?

Table 2: Summary of the covariates

3 Diffusion models

qqs lignes de transition pour expliquer le contenu de la section = modélisation de la vraie position X

3.1 Standard Rotational correlated velocity model (RCVM)

Redit ici qui est X et V par rapport à tes notations précédentes avec y et v. Expliquer que tu considères une extension du modèle (1) pour être anisotrope en A et aussi Σ.

A general formulation for a rotational correlated velocity model is [Albertsen 2018]:

$$\begin{cases} dX(t) = V(t)dt \\ dV(t) = -A(V(t)-\mu)dt + \Sigma dW(t) \end{cases} \quad (3)$$

Toi, tu utilises $\varepsilon_1=\varepsilon_2=\varepsilon$ ou pas ? si tu utilises (1) pourquoi introduire ce modèle (3) ?

where $A = \begin{pmatrix} \frac{1}{\tau_1} & -\omega \\ \omega & \frac{1}{\tau_2} \end{pmatrix}$ and Σ is a lower triangular matrix with positive diagonal elements whereas τ_1 and τ_2 are two autocorrelations time scale for each component of the velocity. This model is indeed a special case of a hypoelliptic diffusion [**Ditlevsen and Samson** 2019]. There is an explicit form for the Gaussian transition density of the velocity process, where the covariance matrix is expressed using a Kronecker sum [**Albertsen** 2018]. ? est-ce utile ?

It is commonly assumed that the movement is isotropic, forcing $\tau_1 = \tau_2 = \tau$, and that the random changes in both components of the velocity are independent with the same variance, in which case $\Sigma = \sigma^2 I_2$ [**Johnson et al.** 2008]. The diffusion parameter σ is written as $\sigma = \frac{2\nu}{\sqrt{\tau\pi}}$, where ν is the unit of ~~homogeneous to~~ a velocity. This allows easier biological interpretation of the parameters in the equation [**Gurarie, Fleming, et al.** 2017]. These simplifications produce the equation 1. → pourquoi avoir généralisé si c'est pour revenir en (1) ?

When parameters τ, μ and ν are constant over time, both the velocity and the position processes can be solved explicitly. The transition density for the velocity process is ~~such that~~ an Ornstein-Uhlenbeck (OU) process with transition density

$$V(t+\Delta)|V(t) = v \sim \mathcal{N}\left((I_2 - \exp(-A\Delta))\mu + \exp(-A\Delta)v, \frac{2\nu^2}{\pi}(1 - e^{-2\frac{\Delta}{\tau}})I_2\right) \tag{4}$$

Note that $\exp(-A\Delta) = e^{\frac{-\Delta}{\tau}} \begin{pmatrix} \cos(\omega\Delta) & \sin(\omega\Delta) \\ -\sin(\omega\Delta) & \cos(\omega\Delta) \end{pmatrix}$ is a rotation matrix of angle $-\omega\Delta$. Intuitively, the next velocity is a weighted mean of the long term mean velocity μ and the previous velocity $v(t)$ rotated by an angle $-\omega\Delta$.

The stationary distribution of the velocity process is $\mathcal{N}\left(\mu, \frac{2\nu^2}{\pi} I_2\right)$. This result can be used to get the long term distribution of the velocity norm and the velocity phase angle. After the velocity process has reached stationarity, that is typically for $t \geq 3\tau$, $||v(t)||$ follows a Rice distribution with mean parameter $||\mu||$ and variance parameter $\frac{2\nu^2}{\pi}$. In the special case $\mu = (0, 0)$, this means that $\mathbb{E}(||v(t)||) = \nu$. Hence, the order of magnitude of the parameter ν is similar to the overall mean empirical velocity norm. The distribution of the phase angle Φ of the velocity has been derived in [**Pawula, Rice, and Roberts** 1982]. When $\mu = 0$, this distribution is uniform over $[-\pi, \pi]$. Hence, taking a look at the histograms of the phase angle of the data can help to assess whether a non zero mean velocity is necessary in the model or not.

By integrating the velocity equation over time one gets the transition density of the location process.

$$X(t+\Delta)|X(t) = x, V(t) = v \sim \mathcal{N}\left(x + \mu\Delta + A^{-1}\left(I_2 - \exp(-A\Delta)\right)(v - \mu), Q(\Delta)\right) \tag{5}$$

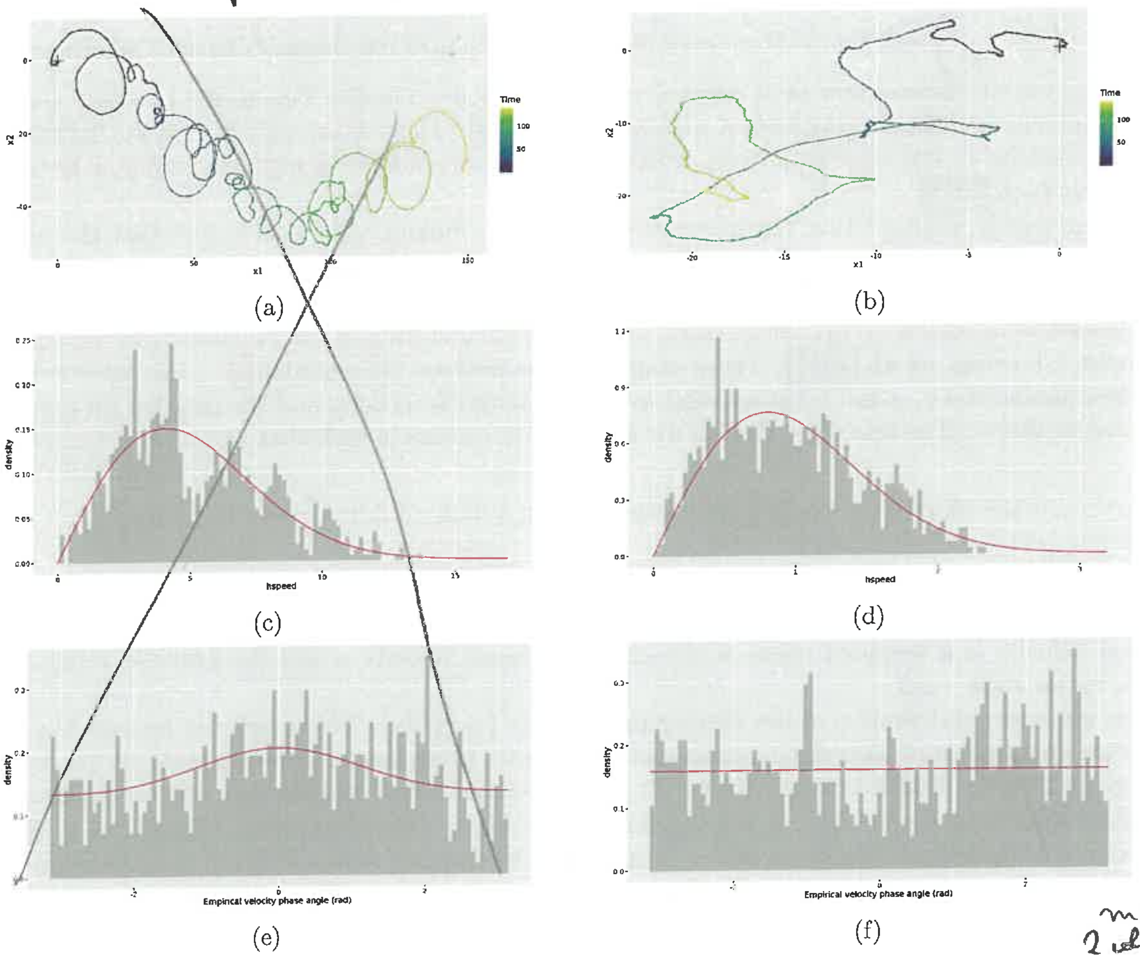
Here, the covariance matrix $Q(\Delta)$ has an explicit form.

$$Q(\Delta) = \left(\Delta - 2\frac{\omega\sin(\omega\Delta) - \frac{1}{\tau}\cos(\omega\Delta)}{\frac{1}{\tau^2} + \omega^2}\exp\left(-\frac{\Delta}{\tau}\right) + \frac{\tau}{2}\left(\frac{\omega^2 - \frac{3}{\tau^2}}{\frac{1}{\tau^2} + \omega^2} - \exp\left(-\frac{2\Delta}{\tau}\right)\right)\right) I_2$$

The position process is not Markovian, but the whole process $(X(t), V(t))^\top$ is. ~~Both~~ The two components of the location and velocity processes are independent. ~~and~~ One can check that for $\omega = 0$, the formulas match the ones of the CTCRW model [**Johnson et al.** 2008]. Figure 7 shows two samples of a RCVM along with the histograms of the empirical velocity norm and phase. Derivation of the formulas given in this paragraph is detailed in appendix A. simulated ?

le modèle contient les observations bruitées ou pas ?

Pas utile ?



est ce utile ?

mettre sur la 2 colonne un exemple de trajectoires avec le nouveau modèle ?

Figure 7: (a)-(b) Two samples of a RCVM with respective parameters $\tau = 5h$, $\nu = 5$ km/h,$\omega = 1$ rad/h, $\mu = (1, 0)$ km/h and $\tau = 5h$, $\nu = 1$ km/h,$\omega = 0$ rad/s, $\mu = (0, 0)$ km/h. The black cross indicates the starting point of the process. (c) - (d) Histograms of the empirical velocities obtained from the samples, along with the theoretical probability density function. (e) - (f) Histograms of the empirical velocity phase angles obtained from the samples, along with the theoretical probability density function.

3.2 Constrained rotational correlated velocity model

new

To get a constrained version of the CVM, i~~t is natural to make use~~ we propose ~~of~~ a new RCVM that includes ~~when the process is close to the shore and the velocity is pointing toward the shore to represent~~ how the animals turn in reaction to the shore. (*) ~~This can be achieved by choosing a~~ smooth function ~~that~~ is zero far away from the shore and non zero close to the shore for ω. Denote $\mathcal{D} \subset \mathbb{R}^2$ the domain of the process X_t = introduire en section 2.2

$$\begin{cases} dX(t) = V(t)dt \\ dV(t) = -A(X(t), V(t))v(t)dt + \frac{2\nu}{\sqrt{\pi\tau}}dW(t) \end{cases}, \quad (6)$$

with ~~Here~~, $A(X(t), V(t)) = \begin{pmatrix} \frac{1}{\tau} & -\omega(D_{shore}(t), \Theta(t)) \\ \omega(D_{shore}(t), \Theta(t)) & \frac{1}{\tau} \end{pmatrix}$ ~~with $D_{shore}(t) = d(X(t), \mathcal{D})$~~, $\Theta(t) =$ tu les as définis avec y... use the equation in section 2.2

~~$(X(t) - p(t), V(t))$ and $p(t)$ is the projection of $X(t)$ on $\partial\mathcal{D}$.~~

Solving explicitly such a model is out of reach due to the non-linearity induced by the distance to the shore and the angle, but it is possible to get an approximate solution. For any $t \geq 0$ and $\Delta > 0$, if Δ is

(*) To do so, we propose to consider ω as a function of $D_{shore}(t)$ and $\Theta(t)$ - This small enough, we can use the approximations

and

$$\forall s \in [t, t+\Delta], D_{shore}(s) \simeq D_{shore}(t) \quad \Theta(s) \simeq \Theta(t)$$

il en faudrait plusieurs ? la version spline de celle que tu proposes ?

In this case, the formulas of section 3.1 are still valid. The question still hanging is : what are the conditions on the function $\omega(D_{shore}, \Theta)$ to ensure that the process is constrained ? We don't formulate a comprehensive answer here, but only give examples of interpretable smooth functions that result in constrained samples of x. An important spatial scale of the rotational movement is $\rho = \frac{\nu}{\omega}$ this is the diameter of rotation. For the process to be constrained, one condition would be that ρ always remains significantly lower than D_{shore}. Moreover, a rotation needs to be considered only for some specific values of the angle Θ, more precisely $\Theta \in [\frac{\pi}{2}, \pi] \cup [-\pi, -\frac{\pi}{2}]$, that is when the narwhal is heading in the direction of the shore. The closer we are to the shore, the faster the velocity should rotate. whereas at a certain distance from the coast, there should be no further effect of the shore, meaning that ω should be 0. Far from the coast, the movement model is simply an integrated Ornstein-Uhlenbeck process, as in the unconstrained case [~~Johnson et al.~~ 2008]. A weight based on the distance to the shore can be defined as follows :

pas le meme qu'avant ?

$$c(D_{shore}) = \exp\left(-\kappa \times \left(\frac{D_{shore}}{D_0}\right)^2\right) \quad \text{for some } D_0 > 0 \text{ and } \kappa > 0 .$$

Based on this weight, a relationship between ω and the angle Θ and the distance to the shore D_{shore} can be formulated as follows :

$$\omega(\theta, D_{shore}) = \frac{\omega_0}{2}\left(\tanh(\lambda(\theta + \frac{\pi}{2})) + \tanh(\lambda(\theta - \frac{\pi}{2}))\right) \times c(D_{shore})$$

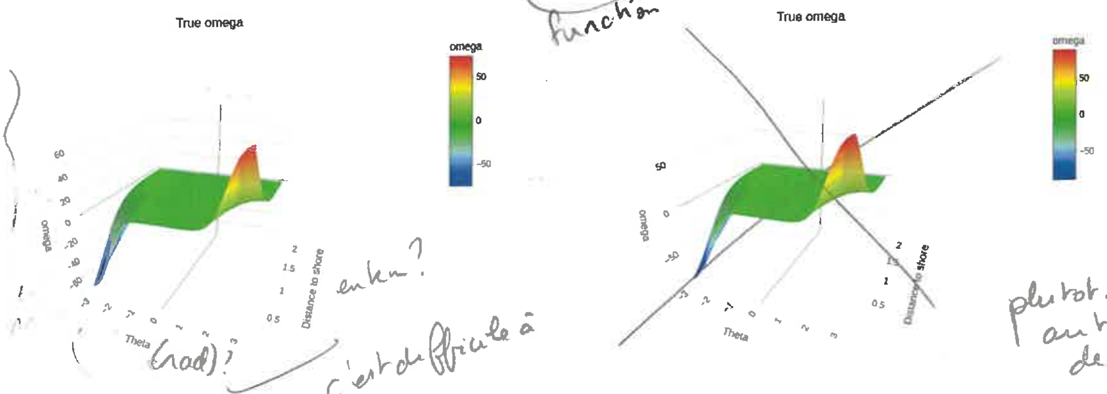
Presenter la version spline

for some $\omega_0 \in \mathbb{R}$, $\lambda > 0$.

This parametrization of ω is only motivated by the observations above, and do not pretend to be optimal. Any smooth function that meets the minimum requirements could be used instead. The general model is very flexible, and different species-specific behaviours could be described. For the narwhals, a tendency to move along the shore could be considered by specifying a function ω that is non zero even when $\Theta \in [\frac{\pi}{2} - \varepsilon, \frac{\pi}{2}] \cup [-\frac{\pi}{2}, -\frac{\pi}{2} + \varepsilon]$, so that when the animal would be on the verge of leaving the shore, its velocity would be tweaked to align with the shoreline. It is also possible to specify τ similarly as a function of D_{shore} and Θ to model specific behaviour near the shore (e.g higher persistence along the shoreline).

Figure 8 shows the smooth function obtained for the parameter ω and different values of D_0, λ and κ.

function



en km ?

(rad) ?

c'est difficile à lire .

plutot mettre un autre exemple de fonction ?

(a) (b)

Figure 8: ω (unit : $rad \cdot h^{-1}$) as a smooth function of Θ and D_{shore}. Values of D_{shore} are in km and values of Θ are in rad. (a) $D_0 = 0.3$ km, $\lambda = 1.5$, $\kappa = 0.5$ and $\omega_0 = \pi$ rad/min. (b) $D_0 = 0.3$ km, $\lambda = 2$, $\kappa = 0.2$ and $\omega_0 = \frac{\pi}{2}$ rad/min.

tu as dit 0.5 dans la definition (l2)

+ Montrer des trajectoires simulées à rajouter dans la fig 7

3.3 Mixed effet model for ship exposure and shore effects

of natural behavior, i.e.,

To assess if the exposure to the ship has ~~a significative~~ an effect on the narwhals movement, a baseline model is first fitted only on the data before exposure, to get an estimation on how the narwhals move when they are not exposed to the disturbance. Then, a response model is fitted on the data during exposure and check whether the estimated smooth parameters deviate significatively from the baseline values. We present here the baseline and response models.

Introduit d'abord les notations des observations

For the baseline model, only the effect of the shore is included through the shore exposure covariate E_{shore} and the angle covariate Θ. plus tard

Although individuals have a tendency to move in the same way, their movement don't have exactly the same characteristics depending on their age and size for instance. This is taken into account by using a mixed effect model for the RCVM parameters. For each narwhal $i \in \{1, \cdots, N\}$, write $t_{exp}^{(i)}$ the first time when it gets in line of sight with the ship (start of exposure). Then for $t \leq t_{exp}^{(i)}$, that is before exposure to the ship, the narwhals motion processes $V^{(i)}(t) = \left(V_1^{(i)}(t) \quad V_2^{(i)}(t)\right)^\top$ and $X^{(i)}(t) = \left(X_1^{(i)}(t) \quad X_2^{(i)}(t)\right)^\top$ are supposed to solve equation 6. Because τ and ν can only be positive, a log link function is used for these parameters. The angular velocity ω is expressed as a smooth function of Θ and D_{shore} through splines. Since both covariates are not on the same scale (one is in *rad* while the other is in *km*), tensor splines were judged more appropriate than thin plate regression splines [Wood 2017]. avant

Données brutes y_{ij} $i = 1, \dots, N$ $j = 1, \dots, n_{i\,pre}$ $n = \sum n_{i\,pre}$

tu mélanges l'introduction du modèle mixte et des données pré- Sépare les concepts

t_{ij} pas réguliers

Puis parle des modèles mixte

dans la section précédente 3.2

Denoting $n_{i,pre}$ the number of observations before exposure for the narwhal i, $\tau^{(i)}$, $\nu^{(i)}$ and $\omega^{(i)}$ the parameters of the diffusion for this narwhal i and $\tau_j^{(i)}$, $\nu_j^{(i)}$, $\omega_j^{(i)}$ their values at time $t_j^{(i)}$ for $j \in \{1, \cdots, n_{i,pre}\}$, the mixed effect model is written Ce n'est pas un modèle mixte - Pour ça il faut rajouter l'équation $y_{ij} = X(t_{ij}) + \varepsilon_{ij}$ et aller sur X avec les effets aléatoires

$$\forall j \in \{1, \cdots, n_{i,pre}\}, \begin{cases} \log(\tau_j^{(i)}) = \log(\tau_{pre}) + b_{\tau,pre}^{(i)} \\ \log(\nu_j^{(i)}) = \log(\nu_{pre}) + b_{\nu,pre}^{(i)} \\ \omega_j^{(i)}(t_j^{(i)}) = \omega_{0,pre} + \sum_{k=1}^{L} \omega_{k,pre}\psi_k(E_{shore}^{(i)}(t_j^{(i)}), \Theta^{(i)}(t_j^{(i)})) + b_{\omega,pre}^{(i)} \end{cases} \tag{7}$$

$\log(\tau_{pre})$, $\log(\nu_{pre})$, $\omega_{0,pre}$ are population intercepts and $b_{pre}^{(i)} = \left(b_{\tau,pre}^{(i)} \quad b_{\nu,pre}^{(i)} \quad b_{\omega,pre}^{(i)}\right)^\top \sim \mathcal{N}\left(0, \operatorname{diag}\left(\cancel{\tfrac{1}{\lambda_{pre}}}\right)\right)$ $\sigma^2_{\omega,pre}$

are the individual parameters, where ~~$\lambda_{pre} = \left(\frac{1}{\sigma^2_{\tau,pre}} \quad \frac{1}{\sigma^2_{\nu,pre}} \quad \frac{1}{\sigma^2_{\omega,pre}}\right)^\top$ with~~ variances $\sigma^2_{\tau,pre}$, $\sigma^2_{\nu,pre}$, $\sigma^2_{\omega,pre}$.

en section 3.2

The ψ_k in equation 7 are known bivariate cubic splines basis functions, constructed through tensor products from the univariate basis functions as described in [Wood 2017], section 4.1.8. $L = q_E \times q_\Theta$ and the numbers of knots in the splines is $q_E - 1$ and $q_\Theta - 1$ for the respective covariates E_{shore} and Θ. It is needed to choose them before hand. One ~~possibility to help choosing the best value is to increase the degree of freedom step by step and find the value for which the decrease in AIC starts to get lower.~~ non tu vas toujours sur-estimer Trend $q_E = q_\Theta = 3$

Equation 7 is based on the approximation that the covariates are constant on each time step $[t_j^{(i)}, t_{j+1}^{(i)}]$, so that the parameters are piecewise constant. ~~which~~ This allows to form a state-space model for estimation as described in section 3.4.

In matrix form, equation 7 can be written as tu ne l'as toujours pas écrit. Il faut vraiment que (7) décrive le modèle entier.

$$\rho_{pre} = X_{pre}\beta_{pre} + Z_{pre}b_{pre} \tag{8}$$

~~where ρ_{pre} is the vector of the parameter values of each time step, that has size $3n$ = number of observations × number of parameters~~ at

$$\rho_{pre} = \left(\log(\tau^{(1)}) \quad \cdots \quad \log(\tau^{(N)}) \quad \log(\nu^{(1)}) \quad \cdots \quad \log(\nu^{(N)}) \quad \omega^{(1)} \quad \cdots \quad \omega^{(N)}\right)^\top$$ is of size $3n_{pre}$

with $\log(\tau^{(i)}) = \left(\log(\tau_1^{(i)}) \quad \cdots \log(\tau_{n_i}^{(i)})\right)$ and similarly for the $\log(\nu^{(i)})$ and $\omega^{(i)}$, X_{pre} and Z_{pre} are respectively the fixed and random effects model matrices with sizes $3n_{pre} \times (L+3)$ and $3n_{pre} \times 3N$, and β is the vector.... Expressions of these matrices are given in Appendix B.

A remettre juste au dessus

$\beta_{pre} = \left(\log(\tau_{pre}) \quad \log(\nu_{pre}) \quad \omega_{0,pre} \quad \omega_{1,pre} \quad \cdots \quad \omega_{q_E q_\Theta,pre}\right)^\top$ is the vector of coefficients to estimate for the pre-exposure model and $b_{pre} = \left(b^{(1)}_{\tau,pre} \quad \cdots \quad b^{(N)}_{\tau,pre} \quad b^{(1)}_{\nu,pre} \quad \cdots \quad b^{(N)}_{\nu,pre} \quad b^{(1)}_{\omega,pre} \quad \cdots \quad b^{(N)}_{\omega,pre}\right)$ is the vector of random effects.

We introduce a second model to analyse data after exposure – Notation $n_{i,post}$ et $n_{post} = \Sigma n_{i,post}$ ça

Then, ~~the~~ response model is designed to assess a deviation from the baseline model due to exposure to the ship. ~~We illustrate this~~ by addding a dependency on the covariate E_{ship} in the parameters τ and ν. For $i \in \{1, \cdots, N\}$, denote $n_{i,post}$ the number of post-exposure observations for the narwhal i.

est ce que tu mets un offset ?

$$\forall j \in \{1, \cdots, n_{i,post}\}, \quad \begin{cases} \log(\tau_j^{(i)}) = \log(\tau_{0,post}) + \sum_{k=1}^{q_\tau} \log(\tau_{k,post})\phi_k(E_{ship}(t_j^{(i)})) + b^{(i)}_{\tau,post} \\ \log(\nu_j^{(i)}) = \log(\nu_{0,post}) + \sum_{k=1}^{q_\nu} \log(\nu_{k,post})\xi_k(E_{ship}(t_j^{(i)})) + b^{(i)}_{\nu,post} \\ \omega_j^{(i)} = \omega_{0,post} + \sum_{k=1}^{q_E \times q_\Theta} \omega_{k,post}\psi_k(D_{shore}(t_j^{(i)}), \Theta(t_j^{(i)})) + b^{(i)}_{\omega,post} \end{cases} \tag{9}$$

Again, $\log(\tau_{0,post})$, $\log(\nu_{0,post})$, $\omega_{0,post}$ are unknown intercepts and $b^{(i)}_{post} \sim \mathcal{N}\left(0, \text{diag}\left(\cancel{\Lambda_{post}}\ \sigma^2\right)\right)$ with ~~$\Lambda_{post} = \left(\frac{1}{\sigma^2_{\tau,post}} \quad \frac{1}{\sigma^2_{\nu,post}} \quad \frac{1}{\sigma^2_{\omega,post}}\right)^\top$~~. q_τ and q_ν are the spline degrees of freedom for the exposure to the ship covariate. We can write the mixed effect models in matrix form with the same notations as above, D

$$\rho_{post} = X_{post}\beta_{post} + Z_{post}b_{post} \tag*{~~(10)~~}$$

Estimation procedure for the vectors of coefficient β_{pre}, β_{post} and the expected values of the random effects b_{pre}, b_{post} given the observations are discussed in section 3.4

3.4 Inference

3.4.1 State space model for the rotational CVM

As discussed in [**Ditlevsen and Samson** 2019], estimation for such hypoelliptic models is challenging. The main difficulties come from

- degenerate noise structure that prevents the use of standard approximation schemes as Euler-Maruyama.
- irregular times of observations : we only get position when the marine mammal reaches the sea surface.
- GPS measurement errors, with gaussian or student distribution.
- partial observations : two processes are involved in equation 1 however only one of them can be observed, that is the position.

plutot en introduction

Several methods already exist to estimate constant parameters τ, ν and μ. If the parameters are constant, it is possible to use a method of moments based on the distributional results about the velocity process [**Gurarie, Fleming, et al.** 2017], but it is only feasible with high frequency data and negligible measurement errors. Maximum likelihood with Kalman filter allows to estimate more efficiently from observations with errors [**Johnson et al.** 2008]. If the parameters are allowed to depend on time through external covariates, then `smoothSDE` package, which is based on [**Johnson et al.** 2008] but exploit the capabilities of both the `mgcv` package [**Wood** 2017] to specify the time dependency through covariates, and the `TMB` [**Kristensen et al.** 2016] package to integrate over the random effect and optimize the likelihood, is a convenient tool for estimation.

du coup tu utilises une methode tres classique je ne suis pas sure qu'il y ait besoin de tout détailler. Ca casse la lecture

Consider equation 1 with parameters τ, ν, μ and ω depending on time. If we suppose these parameters to be constant on each time step $[t_j, t_{j+1}]$ it can be solved explicitly. Denote $\Delta_j = t_{j+1} - t_j$, $v_j = v(t_j)$, A mettre en annexe? ou bien simplifier? ça découle de ce qui est dit en section 3.1 pourquoi le mettre ici?

$x_j = x(t_j)$, $\mu_j = \mu(t_j)$, $\tau_j = \tau(t_j)$, $\nu_j = \nu(t_j)$,$\omega_j = \omega(t_j)$ and $A_j = \begin{pmatrix} \frac{1}{\tau_j} & -\omega_j \\ \omega_j & \frac{1}{\tau_j} \end{pmatrix}$. Then applying Itô's formula gives

$$\forall j \in \{1, \cdots, n\} \quad \begin{cases} v_{j+1} = \mu_j + \exp(-A_j\Delta_j)(v_j - \mu_j) + \zeta_j \\ x_{j+1} = x_j + \mu_j\Delta_j + A_j^{-1}(I_2 - \exp(-A_j\Delta_j))(v_j - \mu_j) + \xi_j \end{cases} \tag{11}$$

where ζ_j and ξ_j are centered normal random variables.

Denote $\alpha_j = \begin{pmatrix} x_{j,1} & x_{j,2} & v_{j,1} & v_{j,2} \end{pmatrix}^\top$ the hidden state and y_j the observed GPS position at time t_j. Measurement errors are supposed to be gaussian in order to be able to use the basic Kalman filter equations. The state space equations are

$$\begin{cases} y_j = Z\alpha_j + \varepsilon_j \\ \alpha_{j+1} = T_j\ \alpha_j + B_j\mu_j + \eta_j \end{cases}$$

il n'y a plus de i ici - c'est dur de suivre tes notations et changements!

where the link matrices are

$$Z = \begin{pmatrix} I_2 & 0_{2,2} \end{pmatrix} \quad T_j = \begin{pmatrix} I_2 & A_j^{-1}(I_2 - \exp(-A_j\Delta_j)) \\ 0_{2,2} & \exp(-A_j\Delta_j) \end{pmatrix} \quad B_j = \begin{pmatrix} \Delta_j I_2 - A_j^{-1}(I_2 - \exp(-A_j\Delta_j)) \\ I_2 - \exp(-A_j\Delta_j) \end{pmatrix}$$

est ce qu'on a besoin de ces nouvelles notations?

$\varepsilon_j \sim \mathcal{N}(0, \sigma_{obs}^2 I_2)$ and the vector $\eta_j = \begin{pmatrix} \xi_{j,1} & \xi_{j,2} & \zeta_{j,1} & \zeta_{j,2} \end{pmatrix}^\top$ is a 4 dimensional normal random vector, with mean zero. Its covariance matrix Q_j can be computed explicitly.

$$Q_j = \begin{pmatrix} M_j & L_j \\ L_j & N_j \end{pmatrix} \tag{12}$$

where M_j, $N_j$and L_j are block matrices of size 2×2 such that

$$M_j = \left(\Delta - 2\frac{\omega_j \sin(\omega_j\Delta_j) - \frac{1}{\tau_j}\cos(\omega_j\Delta_j)}{\frac{1}{\tau_j^2} + \omega_j^2} \exp\left(-\frac{\Delta_j}{\tau_j}\right) + \frac{\tau_j}{2}\left(\frac{\omega_j^2 - \frac{3}{\tau_j^2}}{\frac{1}{\tau_j^2} + \omega_j^2} - \exp\left(-\frac{2\Delta_j}{\tau_j}\right)\right)\right) I_2$$

$$N_j = \frac{2\nu_j^2}{\pi}\left(1 - e^{-\frac{2\Delta_j}{\tau_j}}\right) I_2$$

$$L_j = \begin{pmatrix} l_1 & l_2 \\ l_2 & l_1 \end{pmatrix}$$

c'est comme dans la section 3.1?

where

$$l_1 = \frac{4\nu_j^2}{2\pi\tau_j \times \left(\frac{1}{\tau_j^2} + \omega_j^2\right)} \times \left(1 + \exp\left(-\frac{2\Delta_j}{\tau_j}\right) - 2\exp\left(-\frac{\Delta_j}{\tau_j}\right) - 2\exp\left(-\frac{\Delta_j}{\tau_j}\right)\cos(\omega_j\Delta_j)\right)$$

$$l_2 = \frac{4\nu_j^2}{2\pi\tau_j \times \left(\frac{1}{\tau_j^2} + \omega_j^2\right)} \times \left(\exp\left(-\frac{\Delta_j}{\tau_j}\right)\sin(\omega_j\Delta_j) - \frac{\omega_j\tau_j}{2}\left(1 - \exp\left(-2\frac{\Delta_j}{\tau_j}\right)\right)\right)$$

One can check that for $\omega_j = 0$, the formulas match the ones of the CTCRW model. We coded these formulas in the `smoothSDE` package to allow fast maximum likelihood inference of smooth parameters τ,ν and ω by making the most of the power of the R package `TMB`.

3.4.2 Maximum likelihood estimation

The formulation of a state space model now allows for maximum likelihood estimation using Kalman filter. We detail the maximum likelihood estimation for the response model in section 3.3.

The spline coefficients in equation 9 will be treated as random effects to include a penalization in the log-likelihood. Introduce smoothing parameters $\lambda_\tau, \lambda_\nu, \lambda_{\omega,E}$ and $\lambda_{\omega,\Theta}$. The vectors of spline regression coefficients $\log(\tau_{post})$, $\log(\nu_{post})$, ω_{post}, conditioned on the smoothing parameter are supposed to follow independent multivariate normal distributions with means zero and respective precision matrices $\lambda_\tau S_\tau$, $\lambda_\nu S_\nu$ and $\lambda_{\omega,E} S_{\omega,E} + \lambda_{\omega,\Theta} S_{\omega,\Theta}$. S_τ, S_ν, $S_{\omega,E}$ and $S_{\omega,\Theta}$ are penalization matrices of known coefficients such that $\log(\tau_{post})^\top S_\tau \log(\tau_{post})$, $\log(\nu_{post})^\top S_\nu \log(\nu_{post})$, and $\omega_{post}^\top (S_{\omega,E} + S_{\omega,\Theta})\omega_{post}$ measure the wiggliness of the fitted splines [**Michelot, Glennie, Harris, et al.** 2021]; [**Wood** 2017]. Thus, the chosen prior favours smooth functions over wiggly ones. Now, the mixed effect model equation can be rewritten as $\rho_{post} = X_{post}\beta_{post} + Z_{post} b_{post}$ where the fixed effect coefficients are changed to $\beta_{post} = \begin{pmatrix}\log(\tau_{0,post}) & \log(\nu_{0,post}) & \omega_{0,post}\end{pmatrix}$ and the random effects to

$$b_{post} = \begin{pmatrix}\log(\tau_{post}) & b_{\tau,post} & \log(\nu_{post}) & b_{\nu,post} & \omega_{post} & b_{\omega,post}\end{pmatrix}$$

? [illegible], vraiment étrange

Du coup c'est pas du MLE? expliquer pourquoi tu fais ça ou alors renvoyer à smooth SDE?

The actual random effects $b_{\tau,post}$, $b_{\nu,post}$,$b_{\omega,post}$ also enter the same framework with a "penalization" matrix which is identity and smoothing parameters that are the inverse of the variances to be estimated. All these smoothing parameters are gathered in a vector $\log(\lambda_{post})$ that is estimated from the data.

Denote $y^{(i)}$, $i \in \{1, \cdots, 6\}$, the observed track for the i-th narwhal after exposure, which consists in n_i GPS positions.

The complete likelihood of the data and the random effects is

qu'est ce qui est nouveau ici ? il faut montrer ce qui n'existait pas avant - sinon renvoyer vers les références.

$$L_c(\boldsymbol{b_{post}}, \beta_{post}, \log(\lambda_{post}), \log(\sigma_{obs})) = \prod_{i=1}^{6} p\left(\boldsymbol{b^{(i)}_{post}}, \log(\tau_{post}), \log(\nu_{post}), \omega_{post} | \log(\lambda_{post})\right) \times p\left(y^{(i)} | \boldsymbol{b^{(i)}_{post}}, \log(\tau_{post}), \log(\nu_{post}), \omega_{post}, \beta_{post}, \log(\lambda_{post}), \log(\sigma_{obs})\right)$$

The terms $p\left(b^{(i)}_{post}, \log(\tau_{post}), \log(\nu_{post}), \omega_{post} | \log(\lambda_{post})\right)$ can be computed since $b^{(i)}_{post}$, $\log(\tau_{post})$ and $\log(\nu_{post})$ are independent and their distributions are known multivariate gaussians.

For each individual $i \in \{1, \cdots, 6\}$, the distribution of the hidden state $\alpha_j^{(i)} = \begin{pmatrix} x^{(i)}_{j,1} & x^{(i)}_{j,2} & v^{(i)}_{j,1} & v^{(i)}_{j,2}\end{pmatrix}^\top$ conditioned on previous observations and random effects, i.e $\alpha_j^{(i)} | (y^{(i)}_{1:j-1}, b^{(i)}_{post}, \log(\tau_{post}), \log(\nu_{post}), \omega_{post})$, is $\mathcal{N}(\alpha_j^{(i)-}, R_j^{(i)-})$ where the matrices $\alpha_j^{(i)-}$ and $R_j^{(i)-}$ are computed by the Kalman filter with parameter values according to the conditioned values of the random effects and the equation 9. Hence

bien connu, non ?

$$y_j^{(i)} | (y^{(i)}_{1:j-1}, b^{(i)}_{pre}) \sim \mathcal{N}\left(Z\alpha_j^{(i)-}, Z R_j^{(i)-} Z^\top + \sigma^2_{obs} I_2\right) \tag{13}$$

and the density $p\left(y^{(i)} | \boldsymbol{b^{(i)}_{post}}, \log(\tau_{post}), \log(\nu_{post}), \omega_{post}, \beta_{post}, \log(\lambda), \log(\sigma_{obs})\right)$ is deduced from this result.

Let $\mathcal{L}_c(\boldsymbol{b_{post}}, \beta_{post}, \log(\lambda_{post}), \log(\sigma_{obs}); y) = -\log(L_c(\boldsymbol{b_{post}}, \beta_{post}, \log(\lambda_{post}), \log(\sigma_{obs})))$ be the joint negative log-likelihood of the data and the random effects. Then the likelihood of the data is the integral

$$L(\beta_{post}, \log(\lambda_{post}), \log(\sigma_{obs}); y) = \int_{\mathbb{R}^d} \exp(-\mathcal{L}_c(\boldsymbol{b_{post}}, \theta_{pre}, \log(\lambda_{pre}), \log(\sigma_{obs}); y)) d\boldsymbol{b_{post}}$$

where in this specific case, $d = 3N + q_\tau + q_\nu + L$.

The package smooothSDE relies on the package TMB to optimize this likelihood and get estimations of the coefficients. Laplace's approximation consists in finding a minimum of the function $\mathcal{L}_c$ with respect to b_{post} and then approximating the integral over this parameter by the integral of a gaussian curve. This approximation is then optimized over β_{post}, $\log(\lambda_{post})$ and $\log(\sigma_{obs})$ using gradient methods and auto-differentiation [**Kristensen et al. 2016**]. More precisely, optimization is performed by using the usual optim R function, and plugging in the gradient function calculated by TMB via automatic differentiation. The BFGS method (Broyden–Fletcher–Goldfarb–Shanno algorithm) is generally used for non linear unconstrained problems. If we set constraints on the SDE parameter (maximum values of τ and ν) based on biological expertise for instance, then Limited Memory BFGS with box constraints method should be preferred. In this study, BFGS method was used.
Results of the optimization are estimated parameters intercepts $\hat{\beta}_{post}$, estimated smooothing parameters $\log(\hat{\lambda}_{post})$, estimated modes $\hat{b}^{(i)}_{post}$ of the random effects, and possibly estimated log of the standard deviation $\log(\hat{\sigma}_{obs})$ for the measurement error. Uncertainties on the estimates are derived from the inverse of the hessian matrix of the negative log-likelihood at estimated values and the delta method [**Kristensen et al. 2016**].

dit que le code est disponible sur ...

4 Simulation study

4.1 Simulation of constrained motion

à présenter avant au moment de l'introduction du modèle

In order to test the constrained models we simulated the CRCVM process within different domains for smooth parameters ω depending on the angle Θ and the distance to the boundary of the domain D_{shore}.

~~5~~ Three days long trajectories were sampled in a rectangular domain with sides of length 8 km from the CRCVM with the smooth function for ω displayed in figure 8(a) and other parameters kept constant ($\nu = 4$ km/h, $\tau = 1$ h). For each trajectory, initial velocity was set to 0 and initial position was uniformly sampled in the rectangular domain, at least 500 metres away from the boundary. Time steps were chosen equal to 1 min for high frequency and subsampled to 5 minutes for low frequency. Gaussian measurement errors with low standard deviation $\sigma_{obs} = 3$ m and large standard deviations $\sigma_{obs} = 30$ m were added to provide the actual observations. Of these 100 trajectories, none reach the boundary for $\sigma_{obs} = 3$ m and only one for $\sigma_{obs} = 30$m. However, too small values of D_0 and ω_0, as well as too large measurement errors may produce samples that reach the boundary of the domain. Figure 9 shows some of the 100 samples obtained in the rectangular domain.

donne les valeurs

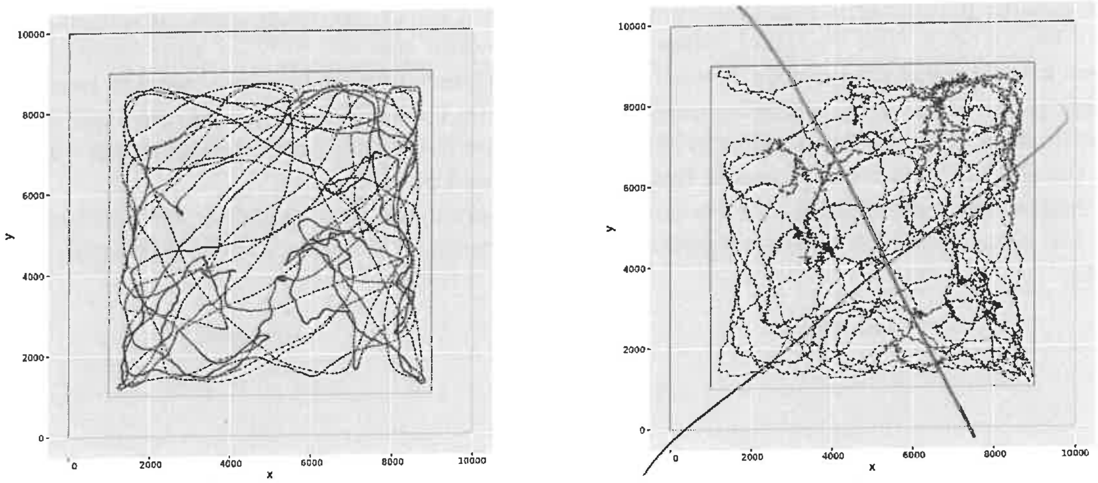
10 trajectories were also simulated in Scoresby Sound fjords system ~~in which the narwhals move under constraints~~. The parameter ω was defined as the smooth function displayed in 8(b). Again, for each trajectory, initial velocity was set to 0 and initial position was uniformly sampled within the sea, with the requirement that it should be at least 500 metres away from the shore. Time steps were also kept constant, equal to 1 min for high frequency and subsampled to 5 min for low frequency. None of the 10 trajectories reached land. Note that if time steps are too large, the assumption of constant covariates between observations is very likely to be violated due to the steepness of the shoreline, resulting in samples that reach land very often. ~~For the fjords system, this already happens if we try to sample directly with 5 min time steps~~. Constrained samples within the geometry of Scoresby sound fjords are shown in Figure 10. Moreover, sampling within the fjords takes much more time than within the rectangular domain, since it is necessary to loop over each polygon of the geometry to compute the nearest shore point for each new position.

et en pratique, ça peut arriver sur les données réelles ?

est ce qu'il y aurait moyen de faire des calculs de distance de façon systématique à l'avance pour accélérer les simus ?

dans les simulations tu peux utiliser un Δ petit et ensuite sous-échantillonner ?

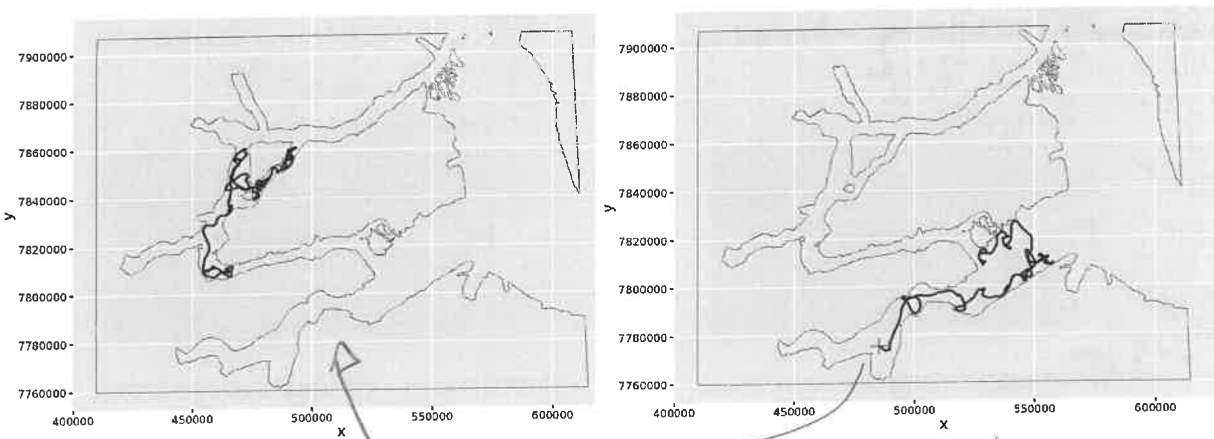
on a pas lecron, en lever la remarque parce que ça ne concerne que les simulations ?



(a) (b)

Figure 9: Samples of the CRCVM. x and y axis are in metres and initial positions are indicated by a red cross (a)-(b) Two samples of time step $h = 1$ min (high frequency) with respective measurement errors standard deviation $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m.

Une seule figure ?



sur le meme graphe

(a) (b)

Figure 10: Samples of the CRCVM. x and y axis are in metres and initial positions are indicated by a red cross. (a)-(b) Two samples of time step $h = 1$ min (high frequency) with respective measurement errors standard deviation $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m.

4.2 Estimation with tensor splines

The samples from previous section were used to estimate the function ω, both in the rectangular domain and in Scoresby Sound fjords system. Covariates E_{shore} and Θ were computed from the observed positions

pas necessaire ?

qu'est ce que ca veut dire ?

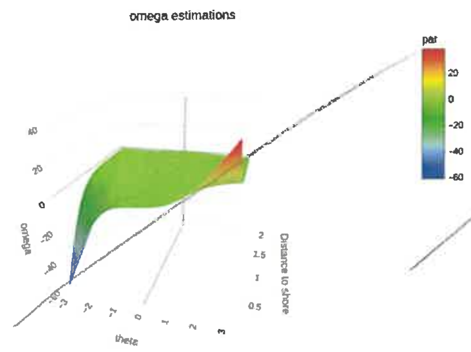
celles simulées ?

ça fait trop ?

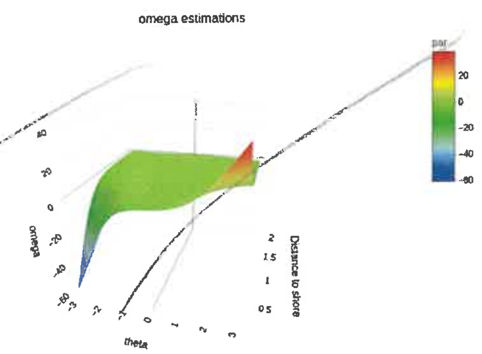
and the smooth parameter ω was estimated using bivariate tensor splines of these covariates, given by the function te in R. The marginal spline degree of freedom was set to ~~5~~ 3 ou for each covariate, which corresponds to 3 knots in the splines. Hence, 25 coefficients (including the intercept) were to be estimated. The other parameters of the diffusion were supposed to be known, as well as the measurement error. Underestimation of the smooth parameter ω is observed close to the shore as Θ approaches $\pm\pi$ (movement toward the shore). This is easily seen in figures 12(a)-(b) and and 14(a)-(b).

As measurement error grows, the covariates that are estimated from the observed positions are becoming less accurate, which makes estimation more challenging. THta is why slight differences in figures 12(a)-(b).

trop difficile à lire -



(a)

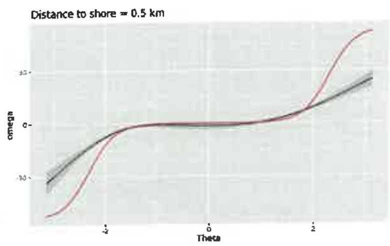


(b)

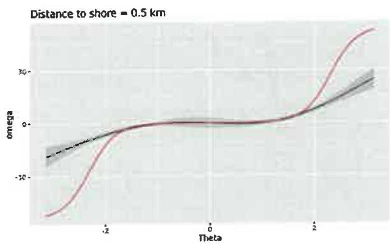
Figure 11: Estimates of the smooth parameter ω as a function of D_{shore} and Θ in the rectangular domain. (a)-(b): Estimation from data with time step $h = 1$ min (high frequency) with respective measurement errors standard deviation $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m.

$\sigma_{obs}=3$ $\sigma_{obs}=30$

plutôt que des figures dans les simulations, mettre les tableaux de biais et RMSE de chaque paramètre estimé

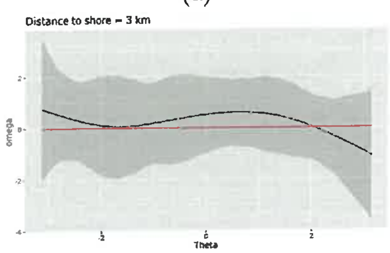


(a)

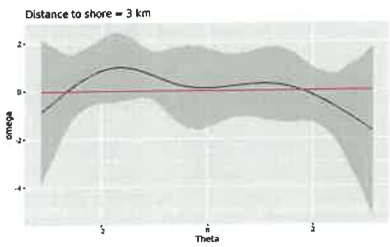


(b)

les CI devraient être + gd. c'est bizarre.



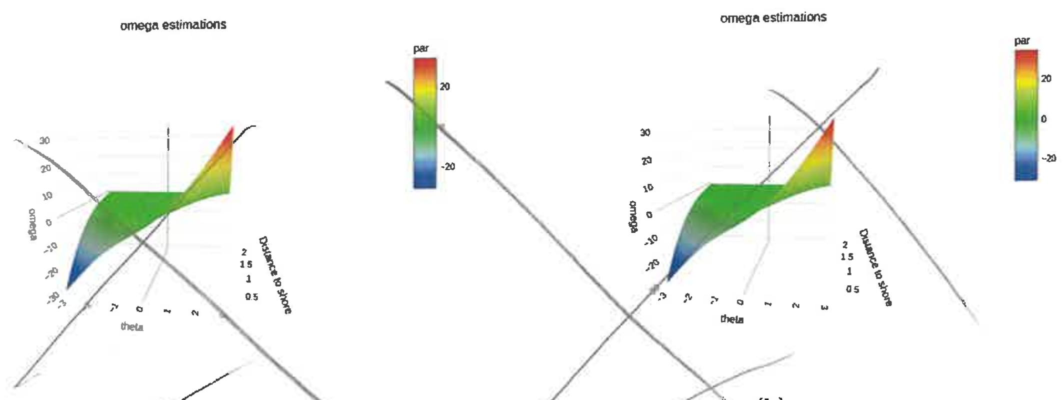
(c)



(d)

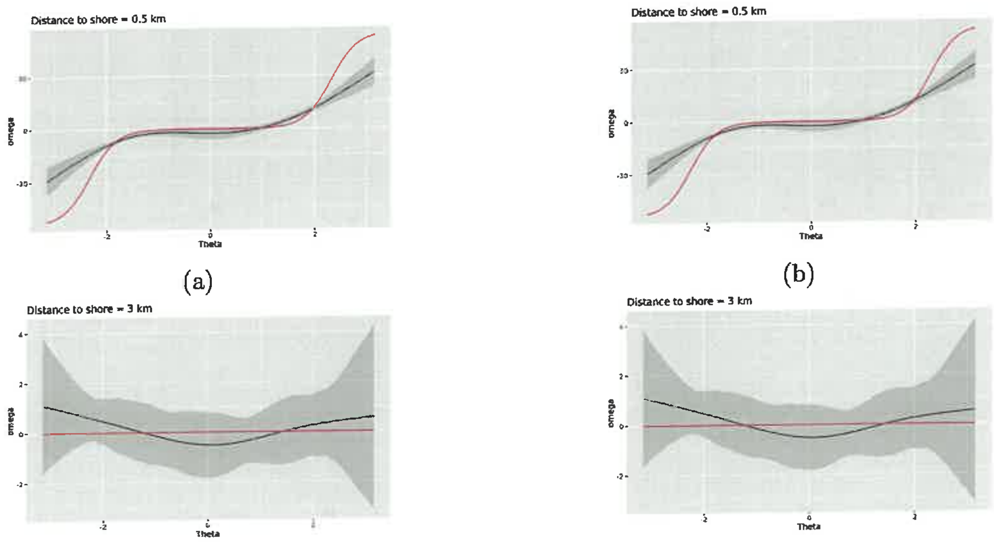
Figure 12: Estimates of the smooth parameter ω as a function of Θ only, in the rectangular domain. The red curve represents the true value of ω. (a)-(b): Fixed distance $D_{shore} = 0.5$ km with respective measurement errors $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m. (c)-(d): Fixed distance $D_{shore} = 3$ km with respective measurement errors $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m.

pas la peine ? garder ces résultats pour ta thèse mais pas le papier



(a) (b)

Figure 13: Estimates of the smooth parameter ω as a function of D_{shore} and Θ in Scoresby Sound fjords system. (a)-(b): Estimation from data with time step $h = 1$ min (high frequency) with respective measurement errors standard deviation $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m.



(a) (b) (c) (d)

Figure 14: Estimates of the smooth parameter ω as a function of Θ only, in Scoresby Sound fjords system. The red curve represents the true value of ω. (a)-(b): Fixed distance $D_{shore} = 0.5$ km with respective measurement errors $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m. (c)-(d): Fixed distance $D_{shore} = 3$ km with respective measurement errors $\sigma_{obs} = 3$ m and $\sigma_{obs} = 30$ m.

5 Estimation of ship exposure effect on real narwhal data

déjà dit?

A baseline model is fitted on the narwhals's track before exposure. It is supposed to capture the characteristics of the movement under normal conditions. Deviation from the baseline is then assessed by fitting a model with covariates E on the tracks during exposure.

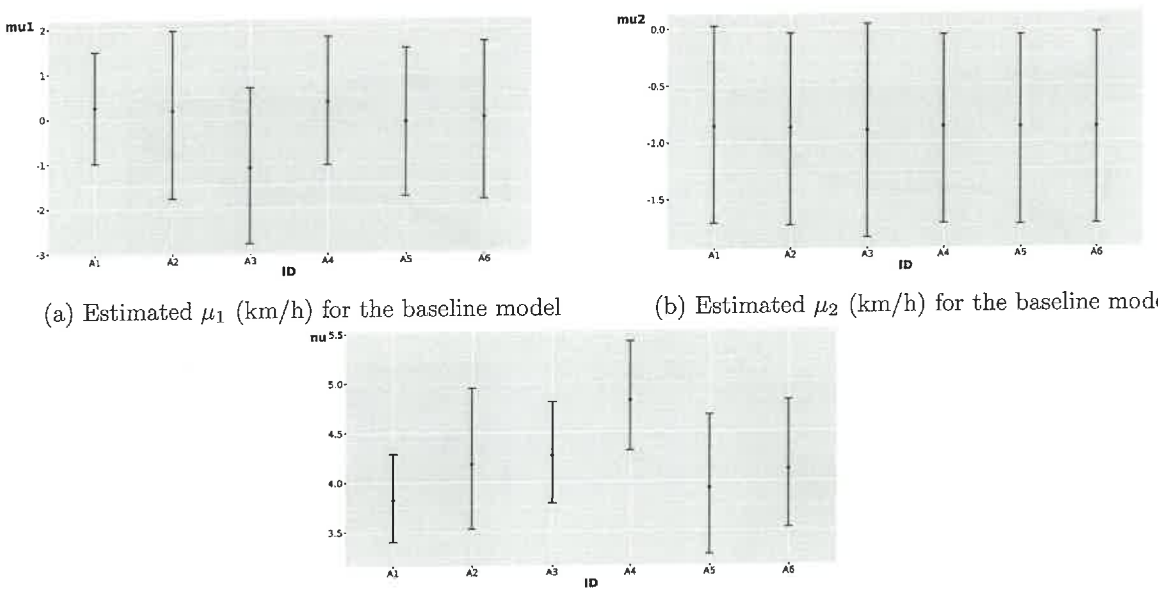
5.1 Estimation without constraints

est ce utile? se concentrer sur le nouveau modèle

First, a simple CVM (see equation ??) was used to assess any behavioural response of the narwhals to the ship exposure.

5.1.1 Baseline estimations

Due to the low variability of the estimates of τ for each individual, it was chosen to only estimate one intercept value for this parameter. It was estimated to $\hat{\tau}_{pre} = 0.90$ km/h with 95% of its values comprised between 0.85 and 1.1 h. In comparison, harbour seal in Alaska were proved to exhibit slightly more persistent motion $\hat{\tau} = 1.51$ h with 95% CI ; $1.30 - 1.75$ **johnson˙continuous˙2008** while bowhead whales in Greenland showed much less persistence $\hat{\tau} = 0.17$ h with 95% CI ; $0.14 - 0.20$ [**Gurarie, Fleming, et al.** 2017]. The intercept value for ν was estimated to be $\hat{\nu}_{pre} = 4.19$ km/h with 95% CI $3.79 - 4.63$. The estimates for the long term velocity μ show that all the narwhals have a tendency to move toward the south ($\mu_2 < 0$) with more individual variability for the east-west direction μ_1. However, confidence intervals are large and forcin μ to $(0, 0)$ is also a reasonable choice.



(a) Estimated μ_1 (km/h) for the baseline model

(b) Estimated μ_2 (km/h) for the baseline model

(c) Estimated ν (km/h)for the baseline model

Figure 15: Estimated parameters for each individual in the baseline model : the y axis are the estimated parameters along with 95% confidence interval and the x-axis represents the narwhals's ID.

The estimated values for ν are all close to 4 km/h with slight variations among individuals. This value should not be come as a surprise since the mean value of the horizontal velocity norm in the data is 4.4 km/h, and it is known that once the velocity process has reached stationarity (typically after 3τ), its norm follows a Rice distribution, whose mean is close to ν when $||\mu||$ is close to 0.

The response model after exposure will be compared to this baseline model to assess any external effect on the narwhals's movement. Examining deviations from this baseline only makes sense when the model is reliable and describes well the narwhals's movement under normal conditions. To check the goodness of fit of the baseline model, two statistics are computed from the observed tracks of each narwhal : mean velocity norm and the mean velocity phase. For the i-th narwhal, the mean velocity norm and phase are approximated as

$$\tilde{v}^{(i)} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i-1} \tilde{v}_j^{(i)} \qquad \tilde{\phi}^{(i)} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i-1} \widehat{(v_j^{(i)}, e)} \text{ where } e = (1,0)$$

where $\tilde{v}_j^{(i)} = \frac{y_{j+1}^{(i)} - y_j^{(i)}}{\Delta_j^{(i)}}$, n_i is the number of measurements, $y_j^{(i)}$ is the GPS position of the narwhal i at time $t_j^{(i)}$ and $\Delta_j^{(i)} = t_{j+1}^{(i)} - t_j^{(i)}$.

Then the intercepts and modes of the random effects of the fitted baseline model are sampled 500 times from the posterior distribution using the joint covariance matrix produced by the optimization of the likelihood. Draws of the parameters of the SDE are deduced from the sampled coefficients and one trajectory is simulated for each narwhal according to the sampled parameters. For each simulated trajectory, mean of the norm and the phase of the velocity are estimated. This produces an histogram for each statistic (norm and phase) and each narwhal. If sampled tracks have enough similarities with the observed tracks, then the observed statistics should appear near the middle of the histograms. This checking procedure was introduced in [**Michelot, Glennie, Thomas, et al. 2022**]. Figures 16 and 17 show respectively the histograms of the mean velocity norm and the mean velocity phase.

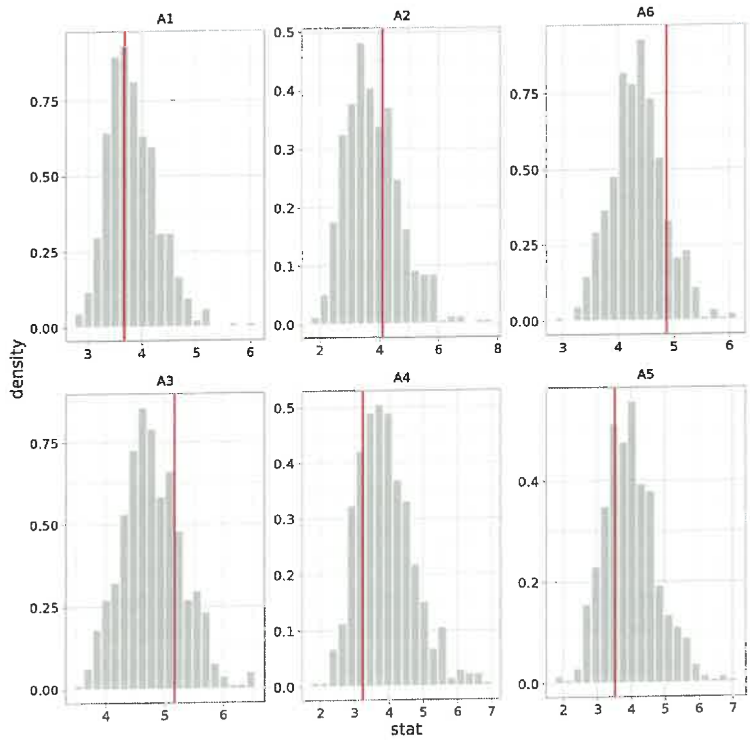


Figure 16: Histogram of the mean velocity computed from trajectories of the fitted model for each narwhal, along with observed values in red.

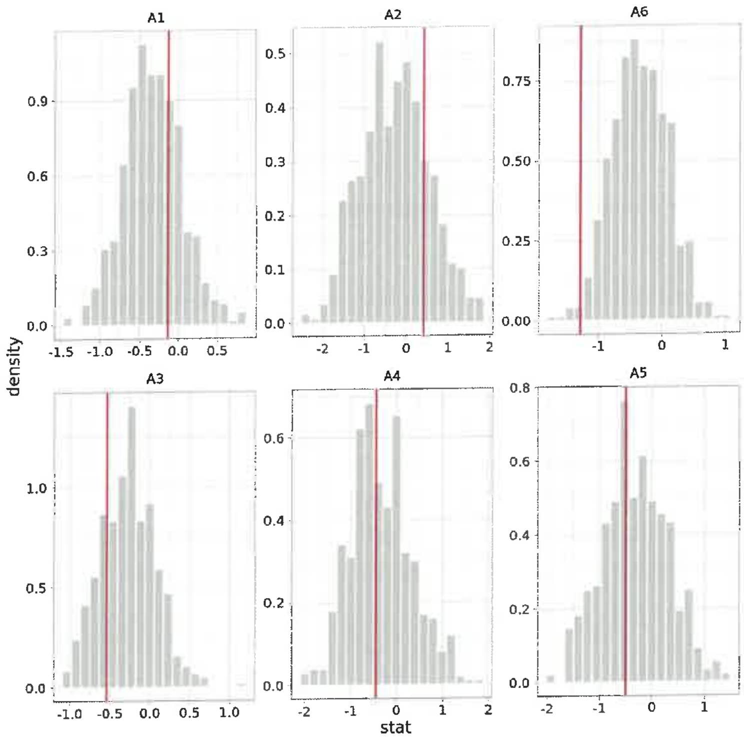


Figure 17: Histogram of the phase angle of the velocity computed from trajectories of the fitted model for each narwhal, along with observed values in red.

The fitted baseline model seems to capture well the mean velocity norm and phase of the observed data for the majority of the narwhals. However, attention should be drawn on the phase angle histogram for the narwhal A6, which reveals a mismatch between the model and the observed phase angle. Narwhal A6 is likely to be heading in the south direction (mean phase angle close to $-\frac{\pi}{2}$).

5.1.2 Response estimations

Once the baseline model is deemed satisfactory, a response model is fitted on the tracks during exposure. We computed the marginal and conditional AIC values with different number of knotrs for the spline estimation of the parameters τ and ν. The marginal AIC is defined in [Wood 2017] as $-2L+2k$ where L is the maximum marginal log-likelihood (of fixed effects), and k is the number of degrees of freedom of the fixed effect component of the model. The conditional AIC also defined in [Wood 2017] is $-2L'+2k'$ where L' is the maximum joint log-likelihood (of fixed and random effects), and k' is the number of effective degrees of freedom of the model. Note that the effective degree of freedom is not properly a number of degrees of freedom, but rather a measure of the flexibility or complexity of the model

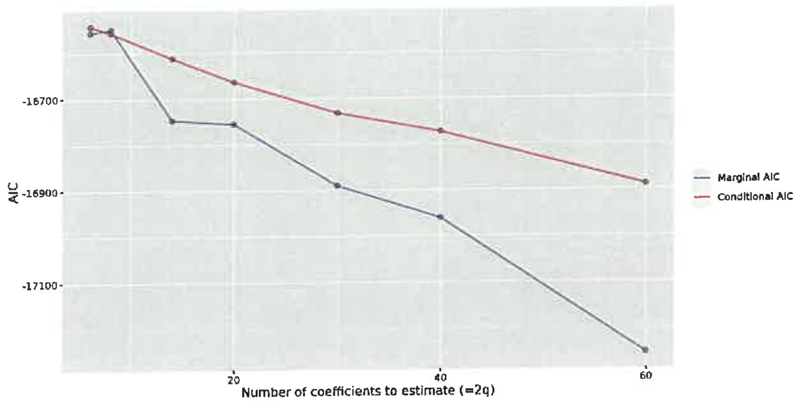
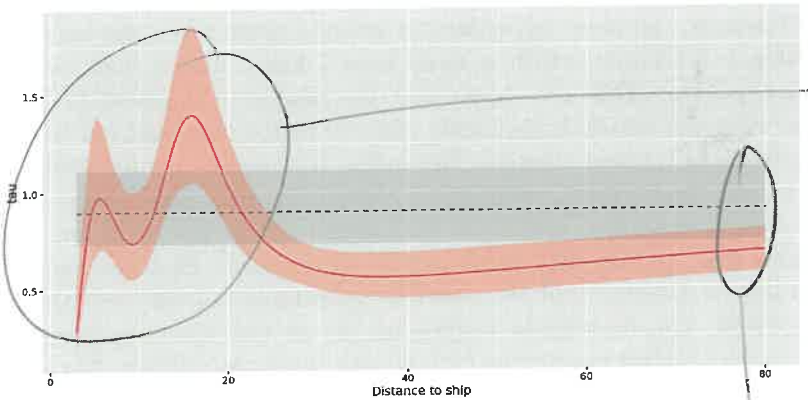


Figure 18

After $q = 7$, the decrease in the marginal AIC curve seems to be lower than it was before. This suggests to choose $q - 1 = 6$ knots for the estimation of the parameters with cubic spline. Note that choosing q a bit lower or higher does not change dramatically the estimated smooth functions. Figure 19 shows the estimated fixed effects of the covariate E for the parameter τ. Figure 20 shows the same plot but for the smooth parameter ν. The variable on the x-axis is the distance to the shore, that is the inverse of E.



probablement trop de degrés dans les splines pour aller chercher artificiellement ces effets là

Figure 19: Estimation of the fixed effects for τ with 95% pointwise confidence interval. The red curve is the estimated function, the dotted line is the intercept estimated in the baseline with its 95% confidence interval in light grey.

il faudrait mettre les effets "pre" comme offset pour revenir à la même valeur

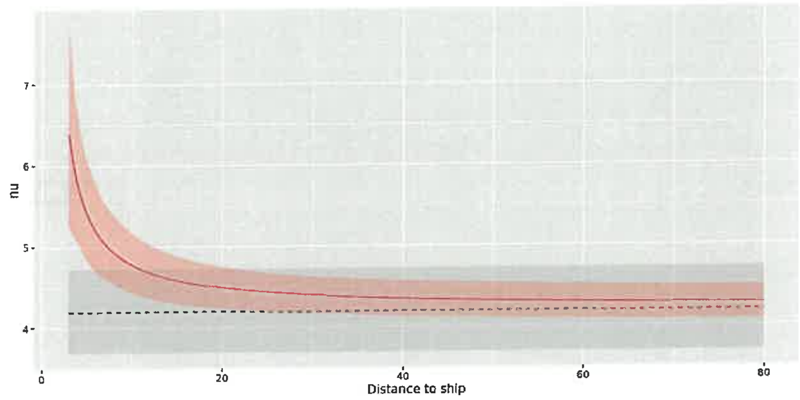


Figure 20: Estimation of the fixed effects for ν with 95% pointwise confidence interval. The red curve is the estimated function, the dotted line is the intercept estimated in the baseline with its 95% confidence interval in light grey.

Deviations from undisturbed behaviour coincide with intervals on the x-axis where the confidence intervals of the baseline estimate and the response estimate are disjoint. This happens for the parameter ν when the narwhals are highly exposed to the ship, more precisely when they are at a distance to the ship lower than 8 km. In this case, ν is significantly higher than the baseline value, proving that narwhals tend to move faster when highly exposed to the ship noise. Again, this should not come as a surprise since, for instance, the mean experimental velocity norm when the narwhals are less than 5 km away from the ship is 5.7 km/h, which is more than 1 km/h higher than the mean experimental velocity norm before exposure. This also shows that the random component in the velocity equation has more importance when the narwhals are highly exposed to the ship, and can be interpreted as more variability in the narwhals horizontal velocity. Regarding the persistence parameter, it is significantly lower than average in two cases. First, when the narwhals are very close to the ship (distance < 3.5 km), it can be up to third times less persistent than usual. Then, up to 25 km away from the ship, no significant deviation from the baseline value can be ascertained. However, more than 25 km away from the ship, it can still be detected that the narwhals persistence is a bit lower than the baseline value (around 0.6 instead of 0.9). This decrease in persistence reveals that the animals gets generally agitated and have a tendency to change their heading more frequently when the military ship is sailing in the fjords.

Narwhals position relative to the ship were considered :

- very close when the distance to the ship was lower than the 5%. quantile, which is 4.04 km.
- close when the distance to the ship was lower than the 25% quantile, which is 9.35 km.
- medium when the distance to the ship was between the 25% and 75% quantile, that is between 9.35 and 25.80 km.
- far otherwise.

Position relative to the ~~coast~~ ship	mean $\hat{\tau}$	mean $\hat{\nu}$	$\frac{\hat{\tau}-\tau_{pre}}{\tau_{pre}}$	$\frac{\hat{\nu}-\nu_{pre}}{\hat{\nu}_{pre}}$
Very close < 4.04 km	0.50	6.1	−44%	+45%
4.04 < Close < 9.35	0.88	5.3	not significaant	+26%
9.35 < Medium < 25.8	1.01	4.6	not significant	not significant
Far > 25.8	0.60	4.3	−33%	not significant

Table 3: Comparison of response and baseline values of the parameters for different levels of exposure to the ship

Voir ce que ça donne avec un offset.

ADD POSTERIOR PREDICTIVE CHECKS

5.2 Estimation with constraints

5.2.1 Baseline estimations

quelle est la valeur de Do finalement ?

The effect of the shore was included directly in the baseline model through bivariate splines of the Distance to the shore and the angle Θ. The mean velocity μ was fixed to $(0, 0)$ and ω was estimated as smooth functions of D_{shore} and Θ.

More details about tensor splines estimations can be found in chapter 4 of [Wood 2017].

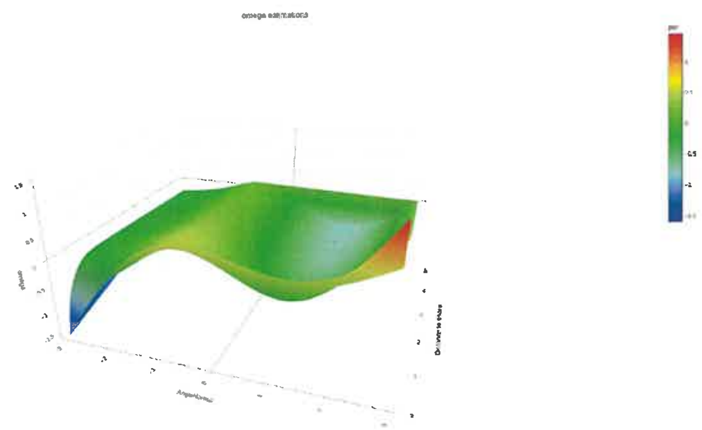


Figure 21: Estimation of smooth parameter ω with tensor splines

5.2.2 Response estimations

6 Discussion and further work

References

Albertsen, Christoffer Moesgaard (June 22, 2018). *Generalizing the first-difference correlated random walk for marine animal movement data.* arXiv: 1806.08582[q-bio]. URL: http://arxiv.org/abs/1806.08582 (visited on 04/14/2024) (cit. on pp. 3, 8, 9).

"Correlation analysis of two-dimensional locomotion paths" (1990). In: *Biological Motion: Proceedings of a Workshop held in Königswinter, Germany, March 16–19, 1989.* Ed. by Wolfgang Alt and Gerhard Hoffmann. Red. by S. Levin. Vol. 89. Lecture Notes in Biomathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 254–268. ISBN: 978-3-540-53520-1 978-3-642-51664-1. DOI: 10.1007/978-3-642-51664-1. URL: https://link.springer.com/10.1007/978-3-642-51664-1 (visited on 04/14/2024) (cit. on p. 3).

Brillinger, David R. (2003). "Simulating constrained animal motion using stochastic differential equations". In: *Institute of Mathematical Statistics Lecture Notes - Monograph Series.* Beachwood, OH: Institute of Mathematical Statistics, pp. 35–48. ISBN: 978-0-940600-55-3. DOI: 10.1214/lnms/1215091656. URL: http://projecteuclid.org/euclid.lnms/1215091656 (visited on 04/14/2024) (cit. on p. 3).

Cioffi, William R et al. (May 29, 2022). *Trade-offs in telemetry tag programming for deep-diving cetaceans: data longevity, resolution, and continuity.* DOI: 10.1101/2022.05.28.493822. URL: http://biorxiv.org/lookup/doi/10.1101/2022.05.28.493822 (visited on 04/14/2024) (cit. on p. 2).

Ditlevsen, Susanne and Adeline Samson (Apr. 1, 2019). "Hypoelliptic Diffusions: Filtering and Inference from Complete and Partial Observations". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81.2, pp. 361–384. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/rssb.12307. URL: https://academic.oup.com/jrsssb/article/81/2/361/7048430 (visited on 04/14/2024) (cit. on pp. 9, 13).

Friedlaender, A. S. et al. (June 2016). "Prey-mediated behavioral responses of feeding blue whales in controlled sound exposure experiments". In: *Ecological Applications* 26.4, pp. 1075–1085. ISSN: 1051-0761, 1939-5582. DOI: 10.1002/15-0783. URL: https://esajournals.onlinelibrary.wiley.com/doi/10.1002/15-0783 (visited on 04/13/2024) (cit. on p. 2).

Garde, Eva et al. (June 1, 2022). "Biological parameters in a declining population of narwhals (*Monodon monoceros*) in Scoresby Sound, Southeast Greenland". In: *Arctic Science* 8.2, pp. 329–348. ISSN: 2368-7460, 2368-7460. DOI: 10.1139/as-2021-0009. URL: https://cdnsciencepub.com/doi/10.1139/as-2021-0009 (visited on 04/14/2024) (cit. on p. 2).

Gurarie, Eliezer, Christen H. Fleming, et al. (Dec. 2017). "Correlated velocity models as a fundamental unit of animal movement: synthesis and applications". In: *Movement Ecology* 5.1, p. 13. ISSN: 2051-3933. DOI: 10.1186/s40462-017-0103-3. URL: http://movementecologyjournal.biomedcentral.com/articles/10.1186/s40462-017-0103-3 (visited on 04/14/2024) (cit. on pp. 3, 4, 9, 13, 20).

Gurarie, Eliezer, Daniel Grünbaum, and Michael T. Nishizaki (June 2011). "Estimating 3D Movements from 2D Observations Using a Continuous Model of Helical Swimming". In: *Bulletin of Mathematical Biology* 73.6, pp. 1358–1377. ISSN: 0092-8240, 1522-9602. DOI: 10.1007/s11538-010-9575-7. URL: http://link.springer.com/10.1007/s11538-010-9575-7 (visited on 04/14/2024) (cit. on p. 3).

Halliday, William D., Matthew K. Pine, and Stephen J. Insley (Dec. 2020). "Underwater noise and Arctic marine mammals: review and policy recommendations". In: *Environmental Reviews* 28.4, pp. 438–448. ISSN: 1181-8700, 1208-6053. DOI: 10.1139/er-2019-0033. URL: https://cdnsciencepub.com/doi/10.1139/er-2019-0033 (visited on 04/13/2024) (cit. on p. 2).

Hanks, Ephraim M., Devin S. Johnson, and Mevin B. Hooten (Sept. 2017). "Reflected Stochastic Differential Equation Models for Constrained Animal Movement". In: *Journal of Agricultural, Biological and Environmental Statistics* 22.3, pp. 353–372. ISSN: 1085-7117, 1537-2693. DOI: 10.1007/s13253-017-0291-8. URL: http://link.springer.com/10.1007/s13253-017-0291-8 (visited on 04/14/2024) (cit. on p. 3).