MASTER BIOSTATISTIQUE

MÉMOIRE DE STAGE

Adéquation et choix de modèle de marche aléatoire des trajectoires des navires de pêches

Auteur :
Rocío JOO
Encadrement :
Sophie BERTRAND
Jean-Michel MARIN

Soutenu le **21 juin 2010** devant le jury composé de :

Christophe ABRAHAM	Professeur, Montpellier SupAgro
Jean-Pierre DAURÈS	Professeur, Université Montpellier I
Gilles DUCHARME	Professeur, Université Montpellier II
Jean-Michel MARIN	Professeur, Université Montpellier II

Remerciements...

Tout d'abord, je remercie profondément Sophie Bertrand pour m'avoir proposé un sujet de stage très intéressant, pour m'avoir accompagné depuis mon stage de licence d'Ingénierie en Statistiques au Pérou, pour supporter mes silences prolongés qui affectent mon travail et sa patience. Merci de m'avoir encourager à suivre le Master Biostatistique. C'est en grande partie grâce à elle et ses efforts pour m'obtenir une bourse, que je suis ici, en France, au sein d'un Master qui me plait.

Je remercie M. Jean-Michel Marin pour s'être proposé de co-encadrer ce stage. Malgré les contraintes de temps et d'espace, ses conseils ont été très importants.

Mes remerciements à l'ensemble de l'IRD, et particulièrement à Philippe Cury, Nicolas Bez, Laurence Vincens, François Gerlotto et Jean-Pierre Lamoureux, pour l'accueil au sein du CRH. A Gérald Hérail et Arnaud Bertrand pour l'accueil à l'IRD Pérou. De la même façon, je remercie l'accueil de l'IMARPE, représenté par Renato Guevara et Miguel Ñiquen.

Je tiens d'une manière générale à remercier de Egide, pour leur financement grâce auquel il m'a été possible de suivre ces études de Master. Pour l'accueil chaleureux et particulièrement celui de Catherine Marzin et Farid Saadoun, pour avoir toujours été disponibles pour répondre à mes questions et levés mes doutes.

Je voudrais remercier très particulièrement Ronan Fablet et Edith Seier pour le soutien dans les moments où je me sentais perdue. Merci beaucoup et muchísimas gracias pour me donner un peu de votre temps précieux.

Muchas gracias también a Daniel, por su placentera y valiosa compañía en las amanecidas. Gracias por la calma y la risa. Gracias también Zayda, por su generosidad al dejarme usar su computadora para sacar mis resultados.

Je tiens à remercier mes camarades du Master Biostatistiques : Rukiye, Jonathan, Jessica, Arnaud, Clara, Mberry, Thomas, Agnès, Asmaa, Héloïse, Aude, Kamel, Senghor et Julie, je les remercie pour leur compagnie, leur aide et leur accueil. Grâce à Thomas et Héloïse je suis tombée amoureuse du LaTeX! Et je remercie encore Thomas : j'ai beaucoup appris lors de nos travaux communs dans le cadre des projets du Master.

Un remerciement spécial à Nicolas (Bez), Pierre (Fréon) et François (Gerlotto). Merci à vous et à vos familles pour m'avoir accueilli au début de mon séjour en France, quand mon français n'était pas encore correct et que j'étais encore perdue dans cette petite ville. Votre chaleureux accompagnement a beaucoup facilité mon adaptation ici.

Résumé

L'étude du mouvement des organismes est essentiel pour la compréhension du fonctionnement des écosystèmes. Dans le cas des écosystèmes marins exploités, cela amène à s'intéresser aux stratégies spatiales des pêcheurs. L'une des approches les plus utilisées pour la modélisation du mouvement des prédateurs supérieurs est la marche aléatoire de Lévy. Une marche aléatoire est un modèle mathématique composé par des déplacements aléatoires. Dans le cas de Lévy, les longueurs des déplacements suivent une loi stable de Lévy. Dans ce cas également, les longueurs, lorsqu'elles tendent vers l'infini (in praxy lorsqu'elles sont grandes, grandes par rapport à la médiane ou au troisième quartile par exemple), suivent une loi puissance caractéristique du type de marche aléatoire de Lévy (Cauchy, Brownien ou strictement Lévy). Dans la pratique, outre que cette propriété est utilisée de façon réciproque sans fondement théorique, les queues de distribution, notion par ailleurs imprécise, sont modélisée par des lois puissances sans que soient discutées la sensibilité des résultats à la définition de la queue de distribution, et la pertinence des tests d'ajustement et des critères de choix de modèle.

Dans ce travail portant sur les déplacements observés de trois bateaux de pêche à l'anchois du Pérou, plusieurs modèles de queues de distribution (log-normal, exponentiel, exponentiel tronqué, puissance et puissance tronqué) ont été comparés ainsi que deux définitions possible de queues de distribution (de la médiane à l'infini ou du troisième quartile à l'infini). Au plan des critères et tests statistiques utilisés, les lois tronquées (exponentielle et puissance) sont apparues les meilleures. Elles intègrent en outre le fait que, dans la pratique, les bateaux ne dépassent pas une certaine limite de longueur de déplacement. Le choix de modèle est apparu sensible au choix du début de la queue de distribution : pour un même bateau, le choix d'un modèle tronqué ou l'autre dépend de l'intervalle des valeurs de la variable sur lequel le modèle est ajusté. Pour finir, nous discutons les implications en écologie des résultats de ce travail.

Mots-clé

Tests d'ajustement, choix de modèle, loi puissance tronquée, loi exponentielle tronquée, mouvement, marche aléatoire de Lévy, trajectoires de pêche.

Abstract

The study of movement of organisms is essential for understanding how ecosystems work. In the case of exploited marine ecosystems, it allows the understanding of spatial strategies of fishermen. One of the most commonly used approaches for modeling top predators movement is the Lévy random walk. A random walk is a mathematical model composed of random steps. In Lévy random walks, the length of the steps follow a stable Lévy distribution. Also Lévy random walks, when lengths of steps tend to infinity (in practice, when lengths of steps are large, say, compared to the median or the third quartile), they follow a power law distribution which characterizes the type of Lévy random walk (Cauchy, Brownian or strictly Lévy). In practice, this property is used on a reciprocal basis without any theorethical foundations. Apart from this fact, tail distributions are modeled by power laws without first discussing important issues as the sensitivity of the results to the definition of tail distribution, the relevance of goodness-of-fit tests and of model selection criteria.

In this work regarding observed movements of three Peruvian anchovy fishing vessels, several models for the tail distribution are compared (lognormal, exponential, truncated exponential, power law and truncated power law) as well as two possible definitions of tail distribution (from the median to the infinite or from

the third quartile to the infinite). In terms of the statistical tests and criteria that were used, truncated distributions (exponential and power law) appeared to be the best. Among other things, they incorporate the fact that, in practice, the boats do not exceed a certain limit on length of steps. Model selection appeared to be sensitive to the choice of the beginning of the tail : for the same boat, choosing one truncated models or the other depends on the range of values of the variable on which the model is adjusted. Finally, we discuss the implications of the results of this work in ecology.

Key-words

Goodness-of-fit tests, model selection, truncated power law distribution, truncated exponential distribution, movement, Lévy random walk, fishing trajectories.

Table des matières

1 Introduction

L'étude du mouvement des organismes est essentielle pour la compréhension du fonctionnement des écosystèmes. Dans le cas des écosystèmes marins exploités, la compréhension des stratégies spatiales des pêcheurs est un facteur clé pour l'étude à différentes échelles spatiales de : l'effort de pêche, les relations prédateur-proie et l'impact des mesures d'administration de la pêche [Bertrand *et al.*, 2007].

La marche aléatoire est une approche statistique aujourd'hui classique pour la modélisation des mouvements des animaux et des humains. Un modèle de marche aléatoire est une formalisation de l'idée intuitive selon laquelle la trajectoire consiste en une succession de pas, chacun de longueur et de direction aléatoire. On peut voir alors la trajectoire comme une séquence discrétisée de déplacements de longueur variable séparés par des changements de cap variables [Bartumeus, 2007].

D'amples débats se donnent dans la littérature au sujet du modèle de marche aléatoire qui serait le plus pertinent pour décrire le mouvement des animaux : modèles de type browniens, browniens corrélés ou marche de Lévy. Il est très probable en fait que le même animal, selon les conditions et les échelles spatio-temporelles auxquelles ses déplacements sont observés, puisse produire des trajectoires cohérentes avec l'un ou l'autre modèle [Benhamou, 2007].

L'une des marches aléatoires les plus utilisées est celle de Lévy, où l'on suppose une loi uniforme pour les changements de cap et une loi stable de Lévy pour les longueurs des déplacements. Dans le cas des marches aléatoires de Lévy (et en dehors du grand débat sur quel serait 'le vrai modèle'), Edwards *et al.* [2007] mettent en question le traitement des données et la méthode d'estimation des paramètres du modèle. En utilisant des poids d'Akaike pour choisir le modèle et des tests G pour sa validation, ces auteurs trouvent que des lois exponentielles sont mieux adaptées que les marches aléatoires de Lévy dans plusieurs exemples en écologie. Le travail de Edwards *et al.* [2007] joue un rôle important dans le débat, dans le cadre du traitement des données [Travis, 2007]. Cependant, nous considérons que la pertinence des tests d'ajustement et des critères de choix de modèle a été fortement négligée dans les discussions.

Dans cette étude nous proposerons plusieurs modèles pour un échantillon de trajectoires des bateaux de pêche à l'anchois du Pérou. Nous utiliserons un sous ensemble de la base des données de [Bertrand *et al.*, 2007]. Nous estimerons les paramètres des modèles par maximum de vraisemblance et nous comparerons les résultats à ceux du papier. En fixant les tests les plus adéquats, nous sélectionnerons le meilleur modèle.

Dans le chapitre suivant on introduira les données de suivi par satellite et le pre-traitement pour récupérer les longueurs des déplacements dans les trajectoires de pêche. Dans le chapitre 3, on donnera la théorie nécessaire pour comprendre les modèles de marche aléatoire de Lévy. On montrera que la loi de la queue de distribution peut caractériser la marche áleatoire de Lévy et ensuite on présentera des lois alternatives pour modéliser la queue de distribution. Dans les chapitres 4 et 5 on présentera les tests d'ajustement et de choix de modèle, respectivement. Les résultats de ces tests seront décrits dans le chapitre 6. Finalement, dans le chapitre 7 on discutera les résultats et leur implications en écologie.

2 Les données

Les données utilisées dans cette étude correspondent au suivi par satellite (VMS : Vessel Monitoring System) des déplacements de trois navires de la pêcherie péruvienne d'anchois (la plus importante pêcherie mono-spécifique au monde) entre 2000 et 2006. Ce système de surveillance par satellite des navires de pêche fournit les enregistrements des coordonnées de position des navires en temps réel et à haute résolution (environ une position par heure). À partir des variables longitude, latitude et temps, on calcule des variables de second et troisième ordre : vitesse, cap, accélération et changement de cap. On utilise un algorithme qui trie les voyages de pêche à partir des critères de distance au port, temps entre enregistrements et vitesse [Bertrand *et al.*, 2005]. On obtient alors des voyages de pêches composés par un minimum de 4 enregistrements VMS. Figure1 montre un exemple de voyage reconnu par l'algorithme.

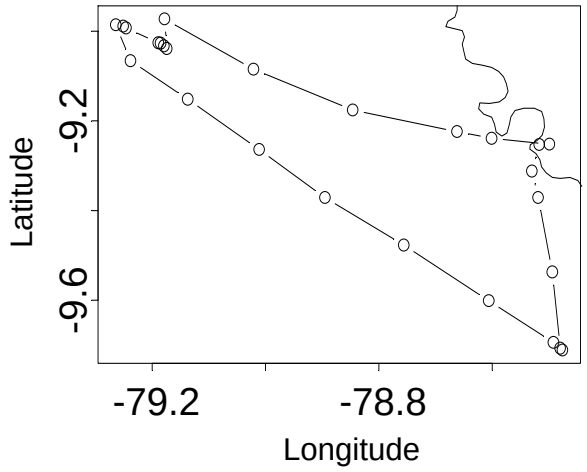


FIG. 1: Voyage de pêche reconnu par l'algorithme proposé par Bertrand *et al.* [2005]

Pour chaque voyage, un deuxième algorithme [Bertrand *et al.*, 2005] divise la trajectoire en déplacements élémentaires dont on calcule les longueurs. Un déplacement [Turchin, 1977] est définie comme l'union des positions consécutives d'un navire sans changement significatif de cap et de vitesse. Sous l'hypothèse que

chaque bateaux (capitaine) a un comportement régulier dans le temps, -c.a.d., sa stratégie entre les voyages de pêche ne varie pas trop -nous grouperons les longueurs des déplacements de tous les voyages réalisés par bateau, chaque bateau étant considéré comme une population statistique distincte.

3 Les modèles de mouvement

Dans ce chapitre nous présenterons les différentes familles de modèles utilisées pour modéliser les mouvements des organismes vivants. Dans la première partie nous décrirons les processus de Lévy qui peuvent se décliner en marches aléatoires de Lévy. On montrera comment la loi de probabilité des queues de distribution des longueurs de déplacements conditionne le type de diffusion associé au mouvement. Plus concrètement, la loi stable de Lévy, loi des longueurs de déplacements du processus de Lévy, tend asymptotiquement vers une loi puissance. Une queue de distribution de loi puissance est décalée vers la droite. Dans la deuxième partie, nous décrirons d'autres lois utilisées dans la littérature pour modéliser les variables décalées vers la droite [Clauset *et al.*, 2007; Newman et Lutcavage, 2005]. On mentionnera aussi d'autres lois qui ont été comparées avec la loi puissance dans le cadre de la modélisation du mouvement. Nous conclurons cette partie par une discussion concernant le choix du début de la queue de distribution à modéliser.

3.1 Processus de Lévy

En avant-propos, nous présentons quelques définitions nécessaires à la compréhension des processus de Lévy [Campillo et Joannides, 2009] :

Definition 3.1 (Processus Stochastique)
Un *processus stochastique* est une famille indexée de variables aléatoires $X = (X_t)_{t\in T}$ définies sur un même espace de probabilité $(\Omega,\mathcal{A},\mathbb{P})$, à valeurs dans un espace mesurable $(E,\mathcal{E})$ et indicées par un paramètre t appartenant à un ensemble T :

$$X : T \times \Omega \to E$$

$$(t, w) \mapsto X_t(w)$$

Dans notre cas, $T \in \mathbb{R}^+$ est un vecteur aléatoire, et $E = \mathbb{R}^2$.

Definition 3.2 (Accroissements de processus aléatoires)
Soit un processus stochastique à temps continu $(X_t)_{t\geqslant 0}$,

1. Il est dit à *accroissements indépendants* lorsque pour tout n et tout $0 \leqslant t_1 \leqslant ... \leqslant t_n < \infty$, les variables $X_{t_1} - X_{t_0}$, $X_{t_2} - X_{t_1}$, ..., $X_{t_n} - X_{t_{n-1}}$ sont mutuellement indépendantes.
2. Il est dit à *accroissements stationnaires* lorsque pour tout h et $t \geqslant 0$, la loi de l'accroissement $X_{t+h} - X_t$ ne dépend pas de t, i.e. $\text{loi}(X_{t+h} - X_t) = \text{loi}(X_h - X_0)$.

Definition 3.3 (Trajectoires d'un processus)
Les trajectoires d'un processus[1] $(X_t)_{t\geqslant 0}$ sont les fonctions :

$$[0, \infty[\ni t \mapsto X_t(w)$$

[1] Dans ce travail, on parlera d'une trajectoire pour faire référence à une représentation (dans le plan) d'un voyage de pêche, comme on l'a vu dans le chapitre 1. À ne pas confondre avec la notion de trajectoire dans un processus stochastique, dont on parlera pour les processus càdlàg.

obtenues à w fixé presque sûrement.

Un *processus* (à trajectoires) *càdlàg* est un processus stochastique tel que pour tout w p.s., les trajectoires $[0,\infty[\ni t \mapsto X_t(w)$ sont **c**ontinues **à d**roite et pourvues de **l**imites **à g**auche.

Les processus à càdlàg, à accroissements indépendants et stationnaires, et issus de 0, sont appelés *processus de Lévy* [2].

La loi d'un *processus de Lévy* est uniquement déterminée par ses lois marginales uni-dimensionnelles P^{L_t}, disons P^{L_1}. Par la propriété d'indépendance et de stationnarité des accroissements, il est clair que P^{L_1} est infiniment divisible. Les *fonctions caractéristiques* de mesures de probabilité infiniment divisibles sont caractérisés par la formule de Lévy-Khintchine :

$$E[e^{iuL_1}] = \exp\left(iub - \frac{c}{2}u^2 + \int (e^{iux} - 1 - iux)\nu(dx)\right) \tag{3.1}$$

où L_1 est la variable aléatoire de longueur de déplacements dans P^{L_1}, $u \in \mathbb{R}$, $c \geqslant 0$ et $\nu(dx)$ est la mesure de Lévy.

Definition 3.4 (triplet Lévy-Khintchine)
Le triplet (b,c,ν) d'une loi infiniment divisible se compose des constantes $b \in \mathbb{R}$ et $c \geqslant 0$ et de la mesure $\nu(dx)$ (appelée mesure de Lévy), qui apparaissent dans l'expression de Lévy-Khintchine de la fonction caractéristique.

Le mouvement Brownien et le processus de Poisson (homogène) sont deux cas particuliers de processus de Lévy. Ils ont des triplets Lévy-Khintchine $(0,c,0)$ et $(0,0,\lambda\delta_1)$, respectivement, où δ_1 est une mesure de Dirac [Applebaum, 2004].

Definition 3.5 (Variable aléatoire stable)
Une variable aléatoire X est dite *stable* si et seulement si $X \overset{loi}{=\!=} dZ + g$, et Z est une variable aléatoire dont la fonction caractéristique est

$$E(e^{iuZ}) = \begin{cases} \exp(-|u|^\alpha[1 - i\beta\tan\frac{\pi\alpha}{2}(\operatorname{sgn} u)]) & \text{si } \alpha \neq 1 \\ \exp(-|u|[1 + i\beta\frac{2}{\pi}(\operatorname{sgn} u)\log|u|]) & \text{si } \alpha = 1 \end{cases} \tag{3.2}$$

où $0 < \alpha \leqslant 2$ est dit *exposant caractéristique* ou *indice de stabilité*, $-1 \leqslant \beta \leqslant 1$ est le *paramètre d'asymétrie*, $d > 0$ est le *paramètre d'échelle* et $g \in \mathbb{R}$ est le *paramètre de localisation* [Nolan, 2010].

Ces distributions sont symétriques autour de 0 quand $\beta = 0$ et $g = 0$:

$$\mathbb{P}(X < y) \equiv \frac{1}{\pi}\int_0^\infty \exp(-d^\alpha \mid u \mid^\alpha)\cos(uy)du \tag{3.3}$$

et dans ce cas la fonction caractéristique de dZ a la forme [Varela *et al.*, 2005]

$$\phi(u) = e^{-d^\alpha|u|^\alpha} \tag{3.4}$$

Les lois stables sont aussi appelées lois alpha stables de Lévy. Pour $\alpha < 2$, les lois stables ont des queues de distribution qui sont asymptotiquement des lois puissance [Nolan, 2010].

Les processus de Lévy sont des processus stochastiques fondés sur la loi (alpha) stable de Lévy, où

[2] Bien que des publications comme celles de Eberlein [2001] et Dubkov *et al.* [2008] utilisent une définition plus générale pour les *processus de Lévy* ; la plupart de la littérature utilise cette définition et nous nous y conformerons

$$\nu(dx) = \frac{C}{|x|^{\alpha+z}}dx \tag{3.5}$$

est la mesure de Lévy avec $0 < \alpha < 2$, $C > 0$ et $z \in \mathbb{R}^d$ (α est l'indice de stabilité) [Applebaum, 2004]. Sauf pour le mouvement Brownien, les variables aléatoires d'un processus de Lévy stable ont une variance infinie et, si $\alpha \leqslant 1$, elles ont aussi une moyenne infinie aussi.

Pour $\alpha < 2$:

$$\lim_{y \to +\infty} \mathbb{P}(X > y) \sim y^{-\alpha} \tag{3.6}$$

ce qui représente une queue de distribution lourde qui suit une loi puissance (c.a.d., qui décroit très lentement) ; contrairement à l'affaiblissement exponentiel de la distribution dans le cas Gaussien [Applebaum, 2004].

3.1.1 Marches aléatoires de Lévy

Les processus de Lévy peuvent être utilisés comme des modèles de marche aléatoire. Un modèle de marche aléatoire est une formalisation de l'idée intuitive qu'une trajectoire consiste en une succession de pas, chacun prenant des valeurs de longueurs et de direction aléatoires. On peut alors voir une trajectoire comme une séquence discrétisée d'événements de déplacement (i.e. **longueurs de déplacements**) séparés par des événements de réorientation successifs (i.e. **changements de cap** - [Bartumeus, 2007]). La loi des longueurs de déplacements et la loi des changements de cap décrivent complètement le processus stochastique. Dans une marche aléatoire de Lévy on suppose une loi uniforme pour les changements de cap et une loi stable de Lévy pour les longueurs de déplacements. À partir de maintenant, on utilisera le résultat montré dans l'équation (3.6) pour caractériser les longueurs de déplacements.

3.1.1.1 Vols et marches de Lévy

Il faut distinguer deux types des marches aléatoires de Lévy : les *vols de Lévy* (VL) et les *marches de Lévy* (ML). Dans les premiers, les "vols" (les déplacements) sont instantanés. Si le système et les échelles étudiés le permettent (e.g. les échelles de temps du processus étudié sont assez grandes par rapport aux temps nécessaires pour accomplir les déplacements unitaires) le processus pourrait être modélisé par un vol de Lévy. Au contraire, si le temps pour effectuer les déplacements unitaire est non négligeable par rapport au temps global de la trajectoire, alors les VL ne pourront pas être utilisés pour décrire le vrai processus dynamique Shlesinger *et al.* [1987] introduisent la marche de Lévy, un processus qui prend en considération le temps d'attente dans la marche aléatoire de Lévy. Les marches de Lévy sont des modifications des vols : elles visitent les mêmes sites et préservent l'auto-similarité spatiale des VL [Shlesinger *et al.*, 1993].

La nature des VL est complètement spécifiée par $p(r)$, probabilité d'un saut de longueur r. Ainsi, la nature des ML est complètement spécifiée par $\Psi(r,t)$ qui est la densité de probabilité de faire un pas de longueur r en un temps t. Cette densité de probabilité est définie par :

$$\begin{aligned} \Psi(r,t) &= \psi(t|r)p(r) \\ \psi(t|r) &= \delta(t - \frac{|r|}{v(r)}) \\ \int drdt\Psi(r,t) &= 1 \\ \lim_{r \to +\infty} p(r) &\sim |r|^{-\alpha} \end{aligned} \tag{3.7}$$

où $\psi(t|r)$ est la densité de probabilité conditionnelle d'un saut de temps t sachant qu'il est de longueur r ; et v est la vitesse qui dépend de la distance de saut r [Shlesinger *et al.*, 1993].

3.1.1.2 Caractérisation des marches aléatoires de Lévy par leur queue de distribution

On peut récrire la troisième équation de (3.7) en utilisant la propriété (3.6) des lois stables de Lévy :

$$\lim_{r \to +\infty} \mathbb{P}(X > r) \sim r^{-\alpha} \tag{3.8}$$

où $0 < \alpha < 2$ [Applebaum, 2004].

Si au lieu de la fonction de survie on utilise la fonction de répartition, alors :

$$\lim_{r \to +\infty} \mathbb{P}(X < r) \sim r^{-\mu} \tag{3.9}$$

où $1 < \mu = \alpha + 1 < 3$ [Benhamou, 2007].

Selon la valeur de μ dans (3.9), qui régit l'affaiblissement des queues de distribution de $p(r)$, on peut caractériser la diffusion du processus [Klafter *et al.*, 1994] :

$$\begin{cases} \mu = 1 & \text{, balistique (loi de Cauchy)} \\ 1 < \mu < 3 & \text{, superdiffusion, } \textit{strictement Lévy} \\ \mu >= 3 & \text{, subdiffusion (mouvement Brownien)} \end{cases}$$

- La diffusion balistique utilise une loi Cauchy avec laquelle l'individu se déplace avec des mouvements droits, sans aucune pause ni changement de direction.
- Le mouvement brownien est un processus de Lévy avec $X_0 = 0$ et $X_t - X_s \sim \mathcal{N}(0, \sigma^2(t-s))$. Il est un processus gaussien centré de noyau de covariance $K(s,t) = s \wedge t$ [Bakry, 2002].

À partir d'ici, et étant clair que les mouvements brownien et balistique sont des cas spécifiques de processus de Lévy, on utilisera pourtant *processus de Lévy* (et *marche de Lévy*) pour désigner des processus (et des marches) strictement Lévy, c.a.d. qui ne sont ni brownien ni balistique, en cohérence avec de nombreux articles faisant référence à ces processus.

Depuis 1996, des nombreuses études modélisant des trajectoires d'animaux et d'humains par des mouvements de Lévy ont été publiés [Atkinson *et al.*, 2002; Austin *et al.*, 2004; Bartumeus *et al.*, 2003, 2005; Benhamou, 2007; Bertrand *et al.*, 2005, 2007; Boyer *et al.*, 2006; Bradshaw *et al.*, 2007; Brockmann *et al.*, 2006; Ramos-Fernández *et al.*, 2004; Rhee *et al.*, 2008; Sims *et al.*, 2008; Viswanathan *et al.*, 1996, 1999, 2000, 2001, 2002]. Edwards *et al.* [2007] montrent cependant qu'en corrigeant les erreurs faites dans plusieurs de ces études, autant dans la manipulation des données que dans l'estimation du paramètre de la loi puissance, le modèle de Lévy n'est plus raisonnable. Cet article est à l'origine d'un grand débat sur la pertinence de l'utilisation des modèles Lévy pour modéliser les déplacements des animaux et sur les conclusions qui peuvent être tirées de ce type d'approches en termes d'écologie [Travis, 2007].

Dans les paragraphes qui suivent, nous présenterons les arguments pour lesquels la marche de Lévy a un intérêt potentiel important en écologie et dans le chapitre suivant nous introduirons quelques méthodes pour valider (ou rejeter) une marche de Lévy et pour choisir un modèle parmi plusieurs modèles.

3.1.1.3 Quelques caractéristiques des marches aléatoires de Lévy et leur importance en écologie

- De façon empirique, une marche aléatoire de Lévy génère des trajectoires qui font intervenir de nombreux 'petits déplacements' intercalés avec de déplacements occasionnels particulièrement longs (Figure 1 à droite). La loi puissance implique que la probabilité d'occurrence d'une longueur de déplacement très grand décroît lorsque cette longueur augmente. Cela en pratique permet d'explorer une surface plus grande et en même temps, de limiter la probabilité de retour à un site déjà visité en comparaison avec une marche brownienne, dont les déplacements particulièrement longs ne sont quasiment pas probables (Figure 1 à gauche).

– Les mouvements de Lévy (sauf pour le mouvement Brownien) n'ont pas de moments à partir du deuxième. Une caractéristique des marches aléatoires de variance infinie est l'invariabilité d'échelle, i.e. la trajectoire peut avoir plusieurs échelles caractéristiques, pourtant aucune d'elle ne domine le processus (pour plus de détail, voir [Shlesinger *et al.*, 1993]. Grâce à cette propriété les points de réorientation ont une distribution fractale. La dimension fractale D de la distribution de ces points est liée linéairement au coefficient μ par [Tsinober, 1994] :

$$D = \mu - 1 \tag{3.10}$$

Si la dimension fractale du processus est faible (avec μ faible), la probabilité d'avoir des déplacements extrêmes (c.a.d. des déplacements très longs ou très courts) augmente. En termes de points de réorientation, les groupes en clusters apparaissent plus étroitement massés et en conséquence plus séparés des groupes de points de réorientation voisins (voir figure 1. Le nombre de niveaux (structures) hiérarchiques dans les groupes de points de réorientation semble augmenter [Mandelbrot, 1977]. Par conséquent, le paramètre μ permet de caractériser le niveau d'agrégation (ou 'patchiness') des points de réorientation ou de changement de direction. Cette agrégation pourrait être causée par le caractère agrégatif des proies. Bertrand *et al.* [2007] proposent d'utiliser le paramètre μ comme indicateur du niveau d'agrégation du poisson et du comportement spatial du pêcheur.

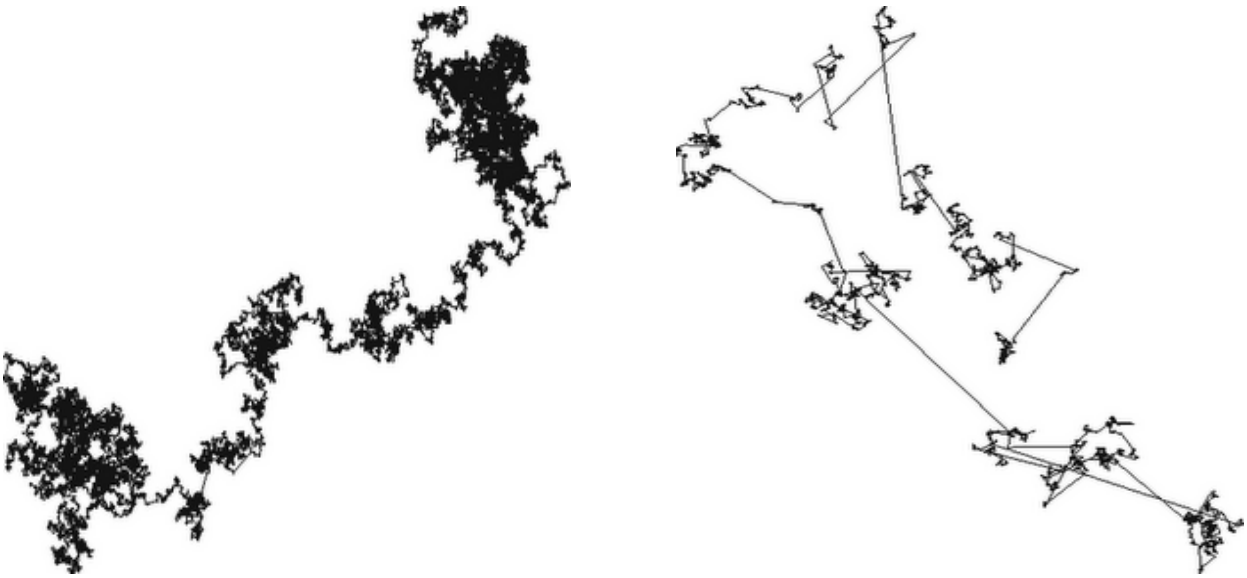


FIG. 1: Marches aléatoires a) Brownien et b) Lévy. Source : http ://politicalcalculations.blogspot.com/2008/12/stock-market-in-lvy-flight.html

3.1.1.4 Vols de Lévy tronqués

Mantegna et Stanley [1994] définissent un vol de Lévy tronqué (TLF) comme un processus stochastique caractérisé par la fonction de répartition :

$$T(x) = \begin{cases} 0 & x > l \\ c_1 L(x) & -l \leqslant x \leqslant l \\ 0 & x < -l \end{cases} \tag{3.11}$$

où

$$L(x) \equiv \frac{1}{\pi} \int_0^{+\infty} \exp(-\mid d \mid^{\alpha} u^{\alpha}) \cos(ux) du \tag{3.12}$$

où u est tel que $\phi(u) = e^{-d^{\alpha}|u|^{\alpha}}$, est la loi stable symétrique de Lévy de l'équation (3.3) d'index $(0 < \alpha \leqslant 2)$ et paramètre d'échelle d $(d > 0)$, c_1 est la constante de normalisation, et l est la borne de la troncature.

La loi de Lévy tronquée est une troncature de la loi de Lévy stable. Nous notons que la fonction de répartition dans l'équation (3.12) est la même que celle de (3.3). La TLF n'est pas stable et a une variance finie. De ce fait, elle converge vers une loi gaussienne, cependant cette convergence est très lente [Varela *et al.*, 2005].

3.2 Lois pour modéliser les queues de distribution des longueurs de déplacements

Dans la section 3.1.1.2, à partir d'une loi stable de Lévy, nous avons montré comment la loi puissance de la queue de distribution peut caractériser le type de marche aléatoire (Cauchy, Lévy ou Brownien). Cependant, ceci est valide uniquement lorsque on part d'une loi stable de Lévy ; c.a.d., il n'y a pas de fondement théorique pour utiliser la loi puissance de la queue de distribution pour caractériser la marche aléatoire si on ne sait pas si la loi des longueurs de déplacements suit une loi stable de Lévy. D'un autre côté, comme nous verrons à continuation, une loi puissance n'est pas définie pour tout $x \geqslant 0$ et a donc besoin d'une troncature inférieure, que nous appellerons 'début de la queue de distribution' pour faire l'analogie avec les marches aléatoires de Lévy (qui ont une queue de distribution qui suit une loi puissance). Nous présentons ici un ensemble de lois de probabilité décalées vers la droite, qui ont été utilisées dans la littérature comme des alternatives à la loi puissance dans plusieurs cas. Ensuite, on évaluera la pertinence de l'utilisation de ces lois pour la modélisation du mouvement.

3.2.1 Lois de probabilité alternatives à la loi puissance

3.2.1.1 Lois puissances tronquées et de troncature exponentielle

La loi puissance (3.9) a la forme suivante :

$$f(x) = Cx^{-\mu} \tag{3.13}$$

où C est une constante.

Comme cette densité de probabilité diverge lorsque $x \to 0$ (3.13), $f(x)$ n'est pas définie pour tout $x \geqslant 0$. De ce fait il est nécessaire de définir une troncature inférieure, a. Avec la contrainte $\mu > 1$, on peut calculer la constante de normalisation C et récrire la densité de probabilité :

$$f(x) = \frac{\mu - 1}{a} \left(\frac{x}{a}\right)^{-\mu} \tag{3.14}$$

Et sa fonction de survie peut s'écrire comme :

$$\mathbb{P}(X > y) = C \int_x^{\infty} f(X) dX = \left(\frac{x}{a}\right)^{-\mu+1} \tag{3.15}$$

Pour les cas où la probabilité de $x \to +\infty$ est trop petite (et non significative), l'utilisation d'une troncature supérieur de X et d'une loi puissance tronquée pour modéliser x seraient plus adéquates.

La loi puissance tronquée a pour densité de probabilité :

$$f(x) = Cx^{-\mu} \tag{3.16}$$

avec $x \in [a, b]$ et $C = \frac{\mu - 1}{a^{1-\mu} - b^{1-\mu}}$.

3.2.1.2 Lois exponentielles non tronquée, tronquée et étirée

Un processus gaussien a une décroissance exponentielle dans sa loi [Ghosal et Roy, 2006], d'où l'intérêt potentiel de modéliser la queue de distribution par une loi exponentielle.

La loi d'une queue de distribution exponentielle peut s'écrire comme

$$f(x) = \lambda e^{-\lambda(x-a)} \tag{3.17}$$

où $x \geqslant a$, a étant le début de la queue de distribution.

Comme dans le cas précédente, pour un X tronqué dans la partie supérieur, la densité de probabilité de la loi exponentielle (tronquée) s'écrit :

$$f(x) = Ae^{-\lambda x} \tag{3.18}$$

où $x \in [a, b]$ et $A = \frac{\lambda}{e^{-\lambda a} - e^{-\lambda b}}$ est la constante de normalisation [Metzler et Klafter, 2002].

Une autre loi utilisée pour modéliser les comportements décalés vers la droite et la loi exponentielle étirée ou loi de Kohlrausch-Williams-Watts (KWW), dont la densité de probabilité s'écrit :

$$f(x) = e^{(x/\tau)^{\beta}} \tag{3.19}$$

avec $0 < \beta < 1$. Quand $\beta = 1$, on récupère la loi exponentielle standard. Le graphique de $f(x)$ versus $log(x)$ est typiquement étiré, d'où le nom de la loi.

Notons que la loi exponentielle étirée a la forme de la fonction caractéristique de la loi stable de Lévy (3.6). Décalée vers la droite, elle est aussi appelée loi complémentaire cumulative de Weibull. En physique, elle est utilisée principalement pour décrire la relaxation dans les systèmes désordonnés ; en biologie, pour la modélisation des événements d'extinction ; en économie, pour les variations des prix dans la bourse des valeurs ; etc. [Laherrere et Sornette, 1998].

3.2.1.3 Loi log-normale

Une autre loi utilisée comme alternative aux lois puissance et exponentielle est la loi log-normale. Une variable aléatoire X est dite suivre une loi log-normale de paramètres μ et σ si une autre variable aléatoire $Y = \log(X)$ suit une loi normale. X admet la densité [Galton, 1879]

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(\log(x) - \mu)^2) \tag{3.20}$$

où $x > 0$.

Quelques caractéristiques

- Cette loi est appelé aussi loi de Galton.
- Elle est décalée vers la droite.
- Le produit de log-normales indépendantes suit une loi log-normale. Sa médiane est égal au produit des médianes de ses facteurs.
- La loi log-normale a été identifiée dans divers domaines comme : géologie (concentration d'éléments et leur radioactivité), médecine humaine (des maladies infectieuses), environnement (distribution de particules, chimiques et organismes dans l'environnement), sciences atmosphériques et aérobiologie (distributions de tailles d'aérosols, de nuages et de paramètres des processus de turbulence), phytomédecine (distribution de la sensibilité aux fongicides dans des populations et distribution des tailles de populations), physiologie des plantes (perméabilité et mobilité des solutés), écologie (abondance des espèces), technologie alimentaire (taille et fréquence de particules de structures), linguistique (nombre de lettres par mot et nombre de

mots par phrase, conversations par téléphone en anglais), sciences sociales et économie (age de mariage, revenu) [Limpert *et al.*, 2001].

3.2.2 Pertinence des différentes lois pour modéliser le mouvement

Nous ferons référence ici aux lois présentées dans la section précédente. Ce sont des lois qu'on pourrait utiliser pour modéliser la queue de distribution des longueurs des déplacements. Dans la littérature, la loi puissance est la plus utilisée pour modéliser le mouvement. On trouve assez souvent la loi exponentielle. Edwards *et al.* [2007] présentent les versions tronquées des lois, plus en phase avec la réalité : les longueurs des déplacements d'un voyage de pêche sont toujours finis. Alors que la loi log-normale est souvent présentée comme une alternative à la loi puissance et exponentielle, elle n'a jamais été utilisée pour modéliser le mouvement. Nous la garderons dans l'ensemble des lois uniquement pour des raisons statistiques (la forme de la loi). Finalement, la loi exponentielle étirée est écartée, tout du moins dans ce travail, pour la difficulté à générer des nombres aléatoires suivant cette loi.

Les modèles de mélange n'ont pas été présentés dans la section 3.2.1. Cette étude vise à la caractérisation du mouvement des bateaux de pêche. La complexité d'un modèle de mélange qui pourrait éventuellement s'ajuster mieux aux données risque de ne pas avoir une interprétation pratique et utile pour notre problème d'application.

En général, les valeurs des paramètres d'une loi et des variables déterminent la vitesse de décroissance de la queue de distribution, son épaisseur et donc les probabilités d'avoir des valeurs plus extrêmes des longueurs de déplacements. Cela est très utile pour caractériser les stratégies de pêche des bateaux (rappelons que pour chaque navire nous avons une population différente de longueurs des déplacements). La figure 2 montre les histogrammes et une trajectoire par bateaux (pris de [Bertrand *et al.*, 2007]).

3.2.3 Choix de la queue des longueurs des déplacements

Dans la théorie du mouvement, il n'existe pas de méthode certaine pour choisir le début de la queue de distribution des longueurs des déplacements. Il est souvent choisi 'à l'oeil', par visualisation de l'histogramme de la variable. Pour être pragmatiques, nous proposons quatre possibles bornes inférieures de la queue de distribution : 1) la médiane, 2) le troisième quartile, 3) le troisième quartile plus 1.5 fois l'écart interquartile et 4) le troisième quartile plus 3 fois l'écart interquartile. La figure 3 montre les queues de distributions pour les différents alternatives de borne inférieure. Nous retenons les deux premières car le nombre des observations groupé par les deux autres est trop faible pour faire de la modélisation.

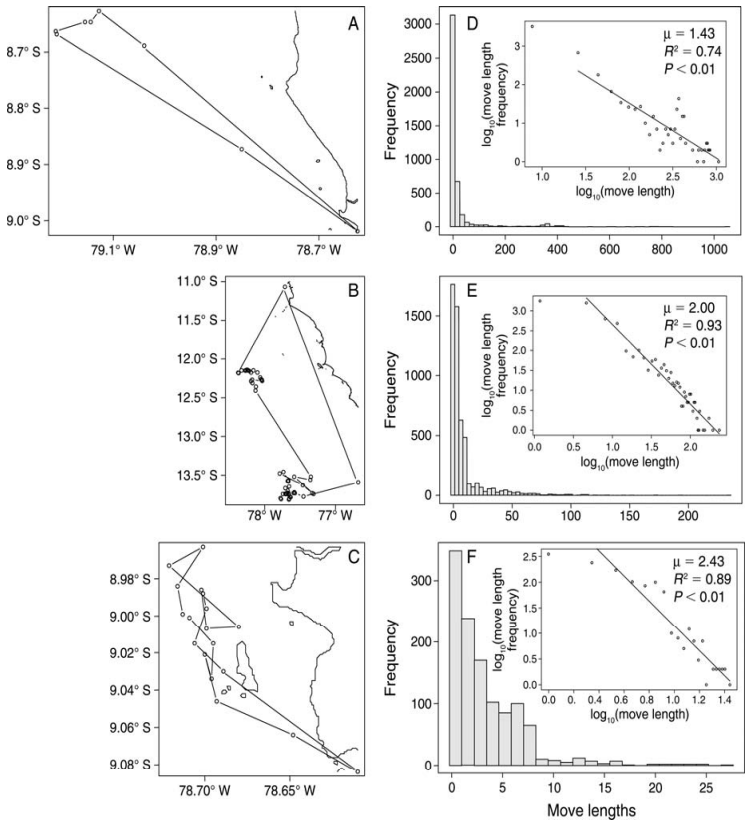


FIG. 2: À gauche : trajectoires des voyages de pêche pour chaque bateau. À droite : histogramme de longueurs de déplacements pour chaque bateau. (Extrait de [Bertrand *et al.*, 2007]).

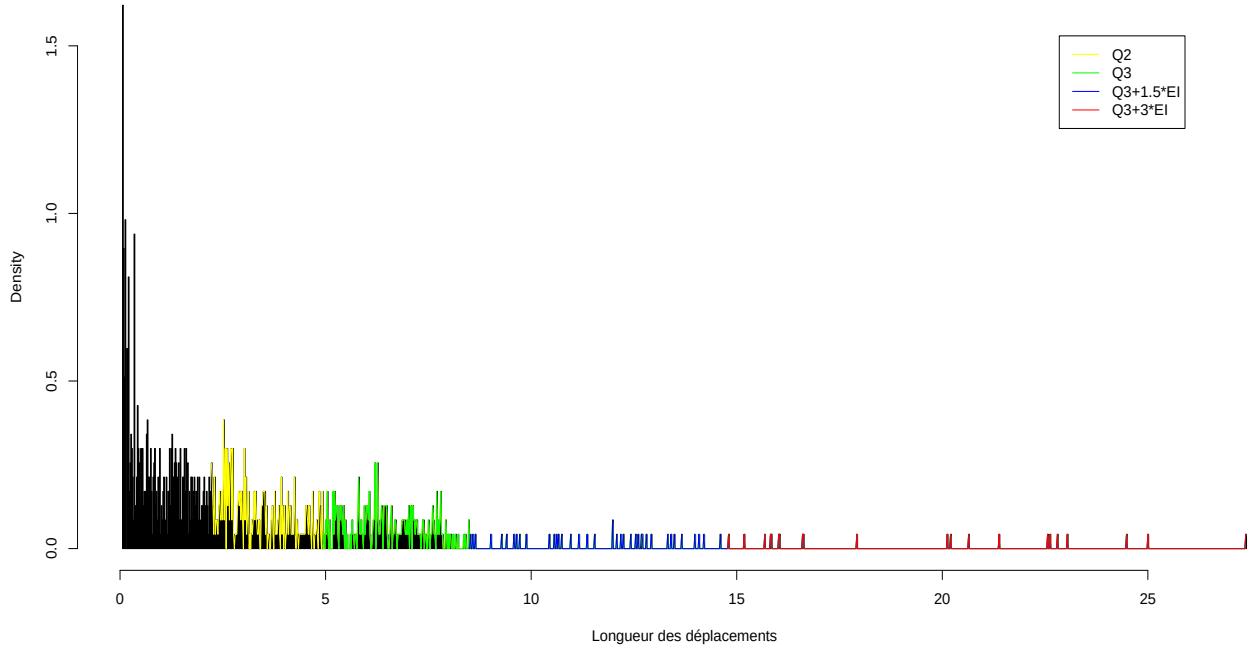


FIG. 3: Histogramme des longueurs des déplacements du bateau 2 et options possibles pour les queues de distribution de l'ensemble des observations. 'Qi' dénote le quartile d'ordre i et 'EI', l'écart interquartile.

4 Tests de qualité d'ajustement

Le meilleur modèle ne s'ajuste pour autant pas nécessairement bien aux données. Pour valider les modèle proposés 3.2.2, nous présentons dans ce chapitre plusieurs tests de qualité d'ajustement ou bien proposés ou bien appliqués aux loi présentés dans la section 3.2.1.

4.1 Analyse Exploratoire

4.1.1 Graphiques pour la loi puissance

Sims *et al.* [2007] procèdent à une revue de la littérature identifiant un comportement Lévy dans les trajectoires des organismes. Les graphiques sont les outils les plus employés pour l'estimation du paramètre μ de la loi puissance. À partir de graphiques log-log des histogrammes de la variable, le paramètre μ est estimé en fonction de la pente d'une régression linéaire du graphique [Bertrand *et al.*, 2005, 2007; Bradshaw *et al.*, 2007; Mårell *et al.*, 2002; Rhee *et al.*, 2008; Viswanathan *et al.*, 1996]. D'après ces travaux, si la courbe qui passe par les points est une ligne droite, alors il y a évidence de loi puissance. Généralement, le coefficient de régression 'r^2' est utilisé comme indicateur de la qualité de l'ajustement. En dehors du biais de la procédure d'estimation [James et Plank, 2007] 'r^2' est uniquement une mesure de la force d'une corrélation linéaire où notre variable d'intérêt a été discrétisé et ensuite transformé ; 'r^2' ne peut pas être considéré comme mesure d'ajustement de la loi à la variable.

4.1.2 Graphiques quantile-quantile (Q-Q)

Un graphique Q-Q est un graphique des quantiles empiriques versus les quantiles théoriques (selon une loi supposée). C'est une technique graphique qui aide à déterminer si la loi proposée donne un ajustement plausible à la variable aléatoire analysée.
La fonction quantile est définie comme

$$Q(p) \coloneqq \inf\{x : F(x) \geqslant p\} \tag{4.1}$$

où $F(x)$ est la fonction de répartition proposée pour X.
On prend comme exemple la loi exponentielle standard :

$$1 - F_1(x) \coloneqq \exp(-x), x > 0 \tag{4.2}$$

dont la fonction de survie est

$$1 - F_\lambda(x) = \exp(-\lambda x) \tag{4.3}$$

Pour savoir si la loi des observations appartient à la famille des lois proposées, paramétrée par $\lambda > 0$, on s'appuie sur les données observées.
Beirlant *et al.* [2004] limitent leur attention à la loi exponentielle. La fonction quantile pour cette loi a la forme

$$Q_\lambda(p) = -\frac{1}{\lambda}\log(1-p) \tag{4.4}$$

où $p \in (0,1)$.

Il y a une relation linéaire simple entre les quantiles d'une loi exponentielle quelconque et les quantiles de l'exponentielle standard correspondante

$$Q_\lambda(p) = \frac{1}{\lambda} Q_1(p) \tag{4.5}$$

où $p \in (0,1)$

À partir d'un ensemble des données $x_1, x_2, ..., x_n$ le praticien remplace la fonction quantile Q de la population par l'approximation empirique $\hat{Q}_n$ définie ci-dessous.

Dans un repère orthogonal, les points des valeurs

$$(-\log(1-p), \hat{Q}_n(p))$$

sont représentés pour différentes valeurs de $p \in (0,1)$. Si le modèle exponentiel permet un ajustement statistique plausible pour l'ensemble des données, alors une ligne droite apparaîtra dans le nuage des points. Lorsque un patron linéaire est obtenu, la pente de la ligne ajustée peut être utilisée comme une estimation du paramètre λ^{-1}.

En effet, si le modèle exponentiel est correct, alors l'équation

$$Q_\lambda(p) = \frac{1}{\lambda}(-\log(1-p)) \tag{4.6}$$

se maintient. On note que l'intercepte doit être 0 puisque $Q_\lambda(0) = 0$.

En général,

$$\hat{Q}_n(p) = x_{i,n}, \tag{4.7}$$

pour $\frac{i-1}{n} < p \leqslant \frac{i}{n}$.

Un choix des valeurs de p très utilisé est donné par

$$p \in \left\{ \frac{1}{n}, 2n, ..., \frac{n-1}{n}, 1 \right\}$$

Un autre choix

$$p \in \left\{ \frac{1-0.5}{n}, \frac{2-0.5}{n}, ..., \frac{n-1-0.5}{n}, \frac{n-0.5}{n} \right\}$$

applique une correction de continuité dont on compare la fonction de discontinuité $\hat{Q}_n$ avec la fonction continue $Q_1(x) := -\log(1-x)$. En outre, ce choix permet d'éviter les problèmes de débordement lorsque $p = 1$. Ceci vaut aussi pour le choix

$$p \in \left\{ \frac{1}{n+1}, \frac{2}{n+1}, ..., \frac{n-1}{n+1}, \frac{n}{n+1} \right\}$$

Une ligne droite peut être ajustée au travers du nuage de points en utilisant un algorithme classique des moindres carrés. La formule des moindres carrés permet l'estimation de la pente a (l'intercepte est 0)

$$\hat{a} = \frac{\sum_{i=1}^n x_{i,n} q_{i,n}}{\sum_{i=1}^n q_{i,n}^2} \tag{4.8}$$

où

$$q_{i,n} := -\log(1-p_{i,n}), i = 1, 2, ..., n$$

Il faut se rappeler que les graphiques sont seulement des aides visuels. Ce qui est déjà très utile, car ils nous permettent d'avoir une première idée de la vérité ou l'erreur de notre hypothèse de départ.

4.2 Tests asymptotiques pour des variables catégoriques : χ^2 et G

Les tests χ^2 et de G sont recommandés pour les cas où la variable mesurée est nominale avec deux ou plus catégories. Le nombre des observations dans chaque catégorie est comparé avec la valeur attendue, obtenue selon la loi supposée pour la variable [Siegel et Castellan, 1988].
Le test χ^2 de [Pearson, 1900] peut s'écrire comme suit :

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.9}$$

où X^2 est le test statistique qui s'approche asymptotiquement à la loi χ^2 (de $n-1$ dégrées de liberté), O_i est une fréquence observée, E_i est une fréquence attendue (théorique) selon l'hypothèse nulle et n est le nombre de catégories.

Le test G ou test des rapports des log vraisemblances [Wilks, 1935] est formulé comme suit :

$$\mathrm{G} = 2\sum_{i=1}^{n} O_i \log\left(\frac{O_i}{E_i}\right) \tag{4.10}$$

D'autres tests semblables sont : le khi-deux modifié, la statistique de Freeman-Turkey, et le test des rapports des log vraisemblances modifié, entre autres [Steele, 2002]. Dans le cas de la loi puissance, les auteurs qui utilisent l'un des deux tests décrits, discrétisent et trient les valeurs observées dans plusieurs catégories. Pour tous ces tests, le résultat risque toujours d'être sensible au choix de la procédure pour le triage. Par ailleurs, en discrétisant une variable on perd beaucoup d'information.

4.3 Tests exacts pour des variables continues

Les tests exacts sont ceux pour lesquels la statistique du test ne converge pas nécessairement vers une loi connue et donc la valeur p est calculée de manière exacte. Nous présentons dans cette section les tests de Kolmogorov-Smirnov, Cramér-von Mises et Anderson-Darling.

4.3.1 Kolmogorov-Smirnov

La statistique de Kolmogorov-Smirnov (KS) pour une fonction de répartition $F(x)$ est

$$D = \sup_x \mid F_n(x) - F(x) \mid \tag{4.11}$$

où la fonction de répartition empirique F_n pour n observations indépendantes et identiquement distribuées est définie comme [Massey, 1951]

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leqslant x} \tag{4.12}$$

Goldstein *et al.* [2004] présentent une nouvelle table pour effectuer le test de KS pour l'ajustement des lois puissances aux données, quand l'exposant de la loi puissance est estimé par maximum de vraisemblance. Clauset *et al.* [2007] utilisent la version de Press *et al.* [1992] de la statistique de KS pour des données non normales :

$$\mathrm{D}* = \max_{x \geqslant \hat{x}_{min}} |F_n(x) - F(x)| \tag{4.13}$$

où $\hat{x}_{min}$ est la valeur de x_{min} qui minimise $D*$.

Le test de Kolmogorov est plus sensible au centre de la distribution que aux extrémités. Goldstein *et al.* [2004] indiquent que si certains paramètres sont estimés à partir des données (comme dans le cas du maximum de vraisemblance), la région critique du KS n'est plus valide.

4.3.2 Anderson-Darling

Pour un échantillon de taille n, la statistique du test d'Anderson-Darling (AD) est :

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x)(1 - F(x))} dF(x) \tag{4.14}$$

où $F_n(x)$ est la fonction de répartition empirique et $F(x)$ est la fonction de répartition théorique. Ainsi que pour le test de Kolmogorov-Smirnov, Bush *et al.* [1983] affirment que les tests d'Anderson-Darling et de Cramér-von Mises sont très conservateurs lorsque les paramètres sont estimés à partir des données. Ils présentent alors une statistique d'Anderson-Darling modifiée, calculée en ordonnant les x_i observés de sorte que $x_1 \leqslant x_2 \leqslant ... \leqslant x_n$ et en construisant

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)[\log(U_{i,n}) + \log(1 - U_{n-i+1,n})] \tag{4.15}$$

où $U_{i,n} = F(x_i)$; $i = 1, 2, ..., n$ [Giles, 2000]. Le test d'AD est une modification du KS qui donne plus de poids aux extrémités.

4.3.3 Cramér-von Mises

La statistique de Cramér-von Mises (CVM) pour tester si les variables aléatoires continues suivent une loi g sur $\mathbb{R}$ est :

$$\omega_n^2 = n \int_{-\infty}^{\infty} \{G_n(x) - G(x)\}^2 dG(x) = n \int_0^1 \{H_n(s) - s\}^2 ds \tag{4.16}$$

où n est la taille de l'échantillon, $H_n(s)$ est la fonction de répartition empirique des variables aléatoires indépendantes $U_1 = G(X_1)$,...$U_n = G(X_n)$, uniformément distribuées dans l'intervalle $[0, 1]$.

Les tests de CVM et AD s'appliquent aux mêmes cas que le test de KS. La différence est que KS repose uniquement sur l'écart maximal entre la distribution empirique est la distribution théorique alors que CVM et AD prennent mieux en compte l'ensemble des données puisque la somme des écarts intervient dans le calcul de la statistique.

Comme dans le cas du test d'Anderson-Darling, ce test est trop conservateur quand les paramètres sont estimés à partir des données [Bush *et al.*, 1983]. Dans ce cas, si $U_{1,n} \leqslant ... \leqslant U_{n,n}$ dénotent les statistiques de rang uniformes $[0, 1]$ correspondantes, ces dernières peuvent être intégrées pour obtenir [Csörgó et Faraway, 1996] :

$$\frac{1}{12n} \leqslant \omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} (U_{i,n} - \frac{2i - 1}{2n})^2 \leqslant \frac{n}{3}, n = 1, 2, ... \tag{4.17}$$

4.4 Tests robustes d'asymétrie et de poids de la queue de distribution

Nous présentons ici les tests robustes de Jarque-Bera, Moors et MC-LR, introduits par Brys *et al.* [2004], Brys *et al.* [2006] et Brys *et al.* [2008], basés sur des mesures d'asymétrie, aplatissement et poids des queues de distribution.

Jarque et Bera [1980] proposent un test de normalité en utilisant les coefficients classiques d'asymétrie et aplatissement :

$$\mathrm{JB} = n\left(\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right) \approx \chi_2^2 \tag{4.18}$$

où b_1 et b_2 sont les estimateurs de l'asymétrie et l'aplatissement, respectivement.
L'asymétrie est définie comme

$$\gamma_1(F) = \frac{\mu_3(F)}{\mu_2(F)^{3/2}} \tag{4.19}$$

est l'aplatissement comme

$$\gamma_2(F) = \frac{\mu_4(F)}{\mu_2(F)^2} \tag{4.20}$$

où F est une fonction de répartition quelconque et les μ_k sont ses moments centraux.
Ce test peut être vu comme un cas spécial de la généralisation suivante : Soit $w = (w_1, w_2, ..., w_k)$ l'estimateur de $\omega = (\omega_1, \omega_2, ..., \omega_k)$, tel que

$$\sqrt{n}(w_1 ... w_k)^t \xrightarrow{loi} \mathcal{N}_k(\omega, \Sigma_k) \tag{4.21}$$

puis, sous H_0, la statistique généralisée T

$$\mathrm{T} = n(w-\omega)^t \Sigma_k^{-1}(w-\omega) \approx \chi_k^2 \tag{4.22}$$

Brys *et al.* [2008] construisent des nouveaux tests d'ajustement analogues à ceux de Brys *et al.* [2004]. Pour le test de **Jarque-Bera**, $k = 2$, $w_1 = b_1$, $w_2 = b_2$, $\omega_1 = \gamma_1$ et $\omega_2 = \gamma 2$. Pour le test de **Moors** [Moors *et al.*, 1996] $w_1 = \frac{F^{-1}(0.75)+F^{-1}(0.25)-2F^{-1}(0.5)}{F^{-1}(0.75)-F^{-1}(0.25)}$ comme mesure robuste d'asymétrie et $w_2 = \frac{F^{-1}(0.875)-F^{-1}(0.625)+F^{-1}(0.375)-F^{-1}(0.125)}{F^{-1}(0.75)-F^{-1}(0.25)}$ comme mesure robuste d'aplatissement. Ce test est résistant au 12.5% des valeurs aberrantes parmi les données. Pour **MC-LR** [Brys *et al.*, 2008] k=3, w_1 = MC, w_2 = LMC et w_3 = RMC, où

$$\mathrm{MC}(F) = \mathrm{med}_{x_1 < m_F < x_2}\, h(x_1, x_2) \tag{4.23}$$

où x_1 ete x_2 sont échantillonnés de F, $m_F = F^{-1}(0.5)$ est h est une fonction kernel

$$h(x_i, x_j) = \frac{(x_j - m_F) - (m_F - x_i)}{x_j - x_i} \tag{4.24}$$

Ainsi

$$\mathrm{LMC}(F) = -\mathrm{MC}(x < m_F) \tag{4.25}$$

et

$$\mathrm{RMC}(F) = \mathrm{MC}(x > m_F) \tag{4.26}$$

Le test MC-LR est résistant à 25% des valeurs aberrantes dans les données.

4.5 Pertinence des tests d'ajustement des lois aux données de longueurs des déplacements

Parmi les tests présentés précédemment, pour notre cas d'étude on écartera ceux qui s'appliquent aux variables catégoriques à cause des désavantages auxquels nous avons fait référence dans la section 4.2 : 1) la sensibilité des résultats au triage des données et 2) la perte d'information en discrétisant les données. Les tests d'Anderson-Darling et de Cramér-von Mises sont préférés à celui de Kolmogorov-Smirnov car les deux

premiers prennent mieux en compte l'ensemble des données que le dernier, qui utilise uniquement l'écart maximal données-modèle. Puisque les paramètres des lois de la section 3.2.2 seront estimés par maximum de vraisemblance, les versions modifiées des tests d'Anderson-Darling et Cramér-von Mises seront utilisées au lieu de la forme classique. Brys *et al.* [2008] montrent que leur test est le plus robuste face aux valeurs aberrantes dans les données en relation à ceux de Jarque-Bera et de Moors. Par conséquent, nous préférerons MC-LR à Jarque-Bera et Moors.

Notre procédure pour le calcul de la statistique T généralisée du test MC-LR (équation 4.22) est la suivante : 1) on calcule la valeur du vecteur des w à partir des données. Pour 'estimer' ω et Σ, on fait plusieurs tirages d'un vecteur de taille égal à notre échantillon observé, suivant la loi testée. Ensuite, on calcule MC, LMC et RMC pour chaque tirage. ω serait le vecteur des moyennes des dernières statistiques et Σ, sa matrice de covariance. Avec cette procédure d'estimation des paramètres, le T calculé ne suit pas un χ^2 (équation 4.22). Pour cette raison, on 'estime' une valeur p. $p(x)$ peut être définie comme la probabilité d'obtenir la valeur de la statistique calculée avec les données ou une valeur plus extrême lorsque l'hypothèse nulle est vraie [Schervish, 1996]. De là, on peut numériquement calculer la valeur de p en générant un grand nombre d'échantillons suivant la loi testée, puis en calculant la statistique (T dans ce cas) pour chaque échantillon et en trouvant dans quelle proportion d'échantillons la valeur de la statistique est plus grande que la statistique calculée pour les données [Seier, 2002]. Cette procédure de calcul de la valeur p sera utilisée aussi pour les tests d'Anderson-Darling et de Cramér-von Mises.

5 Choix de modèle

Différents critères de sélection de modèle ont été proposés en statistique classique. Nous présenterons pour commencer le rapport des vraisemblances, un critère très simple et très utilisé depuis sa préconisation par Neyman et Pearson [1928]. Ensuite, nous introduirons les critères basés sur la divergence de Kullback-Leibler, une distance entre 'le vrai modèle' et le modèle testé. Nous présenterons par ailleurs quelques critères robustes basés principalement sur des mesures de l'erreur d'estimation. Enfin, nous réfléchirons sur les critères à utiliser pour notre cas d'étude.

5.1 Rapport des vraisemblances (LR pour 'likelihood ratio')

Le test de rapport des vraisemblances [Neyman et Pearson, 1928], comme son nom l'indique, est le rapport des vraisemblances maximales des deux modèles emboîtés que l'on veut comparer. Il peut s'écrire comme :

$$\mathrm{LR} = -2(\log(\mathcal{L}_1(\hat{\theta}_j \mid x)) - \log(\mathcal{L}_2(\hat{\theta}_j \mid x))) = -2\log(\frac{\mathcal{L}_1(\hat{\theta}_j \mid x)}{\mathcal{L}_2(\hat{\theta}_j \mid x)}) \tag{5.1}$$

où $\mathcal{L}_j = g_{j,\hat{\theta}_j(x)}$ est la fonction de vraisemblance de $\mathfrak{M}_j$, $g_{j,\hat{\theta}_j(x)}$ est la densité de probabilité de $\mathfrak{M}_j$ proposé, $\hat{\theta}_j$ est l'estimateur de θ_j par maximum de vraisemblance et x représente les données empiriques générées à partir d'une 'vraie loi' f.
Si LR> 1 alors le modèle $\mathfrak{M}_1$ est préféré. Au contraire, si LR< 1 alors $\mathfrak{M}_2$ est choisi.

Vuong [1989] a proposé un test basé sur le rapport des vraisemblances et sur la divergence de Kullback-Leibler (section 5.1) qui, à différence du test LR, peut être utilisé pour des modèles emboîtes et non emboîtes. Il teste l'hypothèse nulle que les deux modèles proposés sont également proche du 'modèle réel' par rapport à l'hypothèse alternative que l'un des modèles est le plus proche. Il ne peut prendre aucune décision si le plus 'proche' modèle est le 'vrai' modèle.

5.2 Critères basés sur la divergence de Kullback-Leibler

Tous les critères visant à identifier le meilleur modèle présentés dans cette section prennent en compte l'ajustement du modèle aux données et, en même temps, la complexité du modèle. Pour préciser mathématiquement le sens du mot 'meilleur', on choisit une mesure de la distance entre 'la vraie densité' f et les densités g_{j,θ_j} de $\mathfrak{M}_j$. On peut définir le 'meilleur' modèle comme étant le $\mathfrak{M}_j$ qui minimise une distance entre f et g_{j,θ_j}. La distance la plus utilisée est la divergence de Kullback-Leibler [Kullback et Leibler, 1951], une mesure non-symétrique des différences entre deux densités de probabilité.

Definition 5.1 (Divergence de Kullback-Leibler)
Soient f et g deux densités de probabilité (par rapport à une mesure dominante η) sur le même X. Alors

$$\mathrm{KL}(f,g) = \int_X f(x)\log(\frac{f(x)}{g(x)})\eta(dx) = \int_X f(x)\log f(x)\eta(dx) - \int_X f(x)\log g(x)\eta(dx) \tag{5.2}$$

avec la propriété $\mathrm{KL}(f,g) \geqslant 0$;

$$\mathrm{KL}(f,g) = 0 \text{ ssi } f = g(\eta - pp) \tag{5.3}$$

KL est dit divergence de Kullback-Leibler [1].

[1] Pour des rappels en théorie de la mesure, voir Annexe.

On définit la distance entre f et les densités $\{g_{j,\theta_j}, \theta_j \in \Theta_j\}$ de $\mathfrak{M}_j$ par

$$\mathrm{KL}(f, \mathfrak{M}_j) = \min_{\theta_j \in \Theta_j} \left(\int_X f(x) \log f(x) \eta(dx) - \int_X f(x) \log g_{j,\theta_j}(x) \eta(dx) \right) \tag{5.4}$$

où le premier terme ($\int_X f(x) \log f(x) \eta(dx)$) est une constante inconnue.
Le meilleur modèle est $\mathfrak{M}_{j*}$ où $j*$ minimise (en j) $\mathrm{KL}(f, \mathfrak{M}_{j*})$, pour $j \geqslant 1$.

- Si $\mathrm{KL}(f, \mathfrak{M}_{j*}) = 0$ (cas 1), $f = g_{j*,\theta_{j*}}$ pour un certain $(j*, \theta_{j*})$. θ_{j*} est la 'vraie' valeur du paramètre et $\mathfrak{M}_{j*}$ est le 'vrai' (et meilleur) modèle pour X (c.a.d., $X \sim g_{j*,\theta_{j*}}$).
- Si $\mathrm{KL}(f, \mathfrak{M}_{j*}) > 0$ (cas 2), θ_{j*} est la 'meilleure' valeur du paramètre et $\mathfrak{M}_{j*}$ est le 'meilleur' modèle pour X (c.a.d., $g_{j*,\theta_{j*}}$ est la densité la plus proche de f selon KL).

5.2.1 Critère d'Information de Takeuchi (TIC)

En minimisant une approximation de la distance de Kullback-Leibler de l'équation (5.4) et en estimant θ par maximum de vraisemblance [Ducharme, 2009], on trouve le critère d'information introduit par Takeuchi [1976] :

$$\mathrm{TIC}_j = -2 \log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + 2 \operatorname{tr}[\widehat{J}(\hat{\theta}_j) \widehat{I}(\hat{\theta}_j)^{-1}] \tag{5.5}$$

où TIC_j est le TIC du modèle $\mathfrak{M}_j$, $\mathcal{L}_j = g_{j,\hat{\theta}_j(x)}$ est sa fonction de vraisemblance, $g_{j,\hat{\theta}_j(x)}$ est la densité de probabilité de $\mathfrak{M}_j$ proposé, $\hat{\theta}_j$ est l'estimateur de θ_j par maximum de vraisemblance, x représente les données empiriques générées à partir d'une vraie loi f, $\widehat{J}(\hat{\theta}_j)$ et $\widehat{I}(\hat{\theta}_j)$ sont les deux des matrices $K \times K$ (étant K la dimension de θ) telles que

$$\widehat{J}(\hat{\theta}_j) = \left[\frac{\partial \log(g(x_i \mid \hat{\theta}_j))}{\partial \theta_j} \right] \left[\frac{\partial \log(g(x_i \mid \hat{\theta}_j))}{\partial \theta_j} \right]^T \tag{5.6}$$

est l'estimateur de

$$J(\theta_j) = E_f \left[\left(\frac{\partial \log(g(x \mid \theta_j))}{\partial \theta_j} \right) \left(\frac{\partial \log(g(x \mid \theta_j))}{\partial \theta_j} \right)^T \right] \tag{5.7}$$

et

$$\widehat{I}(\hat{\theta}_j) = -\frac{\partial^2 \log(g(x \mid \hat{\theta}_j))}{\partial \theta_j \theta_{j'}} \tag{5.8}$$

est l'estimateur de

$$I(\theta_j) = E_f \left[-\frac{\partial^2 \log(g(x \mid \theta_j))}{\partial \theta_j \theta_{j'}} \right] \tag{5.9}$$

pour un certain $\theta_j \in \Theta_j$.
Le modèle avec le TIC_j minimum sera retenu.

Quelques remarques
- Le TIC n'impose pas l'appartenance du vrai modèle à la famille des modèles ajustés.
- Il est valide pour les modèles emboîtés et non emboîtés.
- Le premier terme est la log-vraisemblance estimée au point x : elle augmente avec la qualité de l'ajustement du modèle à la donnée.
- Le deuxième terme pénalise les modèles trop paramétrés.
- Si le 'meilleur' modèle est en réalité le 'vrai' modèle, alors TIC est équivalent à AIC (le critère d'information d'Akaike, section 5.2.2)

– Puisque le deuxième terme ne dépend pas de la taille de l'échantillon n, ce terme reste stable.
– Par contre, avec un n très grand le premier terme augmente et il risque de dominer TIC si le deuxième terme reste trop petit ; c.a.d., pour un nombre des paramètres fixé, plus les observations sont nombreuses, moins la pénalité est importante.

5.2.2 Critère d'Information d'Akaike (AIC)

Si $g_{j,\theta_j} = f$, c.a.d., $\mathfrak{M}_{j*}$ est le 'vrai modèle', l'équation (5.5) peut se récrire comme :

$$-2\log(g_{j,\hat{\theta}_j(x)}(x)) + 2\dim(\theta_j) = -2\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + 2K_j = \mathrm{AIC}_j \tag{5.10}$$

où AIC_j est le critère d'information d'Akaike [Akaike, 1973] de $\mathfrak{M}_j$, $\mathcal{L}_j = g_{j,\hat{\theta}_j(x)}$ est la fonction de vraisemblance, $g_{j,\hat{\theta}_j(x)}$ est la densité de probabilité de $\mathfrak{M}_j$ proposé, $\hat{\theta}_j$ est l'estimateur de θ_j par maximum de vraisemblance, x représente les données empiriques générées à partir d'une 'vraie' loi f, et K_j est la dimension de θ_j.
Le modèle dont le AIC_j est le plus petit sera préféré.

AIC pour des petits échantillons L'equation (5.10) est basée sur des approximations asymptotiques et n'est valable que pour les échantillons suffisamment grands. Le AIC corrigé pour des petit échantillons (AIC_c) est introduit par Sugiura [1978]. Il peut s'écrire comme :

$$\mathrm{AICc}_j = -2\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + 2K_j + \frac{2K_j(K_j+1)}{n-K_j-1} \tag{5.11}$$

où AIC_j est le AIC de $\mathfrak{M}_j$ et K_j est la dimension de θ_j et n est la taille de l'échantillon. Comme le deuxième terme de l'équation, le troisième terme est un terme de correction du biais (de deuxième ordre). Burnham et Anderson [2004] conseillent son utilisation lorsque $n/k < 40$. Dans la pratique, quand n devient grand, AIC_c converge vers AIC.

Poids d'Akaike Burnham et Anderson [2004] introduisent les poids d'Akaike :

$$w_j(\mathrm{AIC}) = \frac{\exp\{-\frac{1}{2}\triangle_j(\mathrm{AIC})\}}{\sum_{i=1}^{J}\exp\{-\frac{1}{2}\triangle_i(\mathrm{AIC})\}} \tag{5.12}$$

où $\triangle_j(\mathrm{AIC}) = \mathrm{AIC}_j - \min \mathrm{AIC}$ et J est le nombre des modèles testés. $\triangle_j(\mathrm{AIC})$ peut s'interpréter comme la perte d'information quand on ajuste un modèle $\mathfrak{M}_j$ au lieu de $\mathfrak{M}_{min}$. L'expression du numérateur, $\exp\{-\frac{1}{2}\triangle_j(\mathrm{AIC})\}$, fournit la vraisemblance du modèle [Akaike, 1981] à partir des données : $\mathcal{L}_j(\hat{\theta}_j \mid x)$.

La vraisemblance relative du modèle $\mathfrak{M}_j$ versus $\mathfrak{M}_i$ est $\mathcal{L}_j(\hat{\theta}_j \mid x)/\mathcal{L}_i(\hat{\theta}_i \mid x)$ et est appelée *ratio d'évidence*. Les ratio peuvent avoir des valeurs comprises entre 0 et ∞. Il est plus pratique de normaliser les vraisemblances des modèles de telle façon qu'elles somment 1 et de les traiter comme des probabilités. Ces probabilités sont les poids d'Akaike de l'équation (5.12). Le AIC_j dont le poids est le plus grand sera préféré.

Quelques remarques
– L'AIC requiert l'appartenance du vrai modèle à la famille des modèles ajustés.
– Il est valide pour les modèles emboîtés et non emboîtés.
– Le premier terme des équations (5.10) et (5.11) est la log-vraisemblance estimée au point x : elle augmente avec la qualité de l'ajustement du modèle à la donnée.
– Le deuxième terme des équations (5.10) et (5.11) pénalise les modèles trop paramétrés.
– Le troisième terme de l'équation (5.11) corrige le biais pour les petits échantillons.
– La correction exacte du biais dépend de la loi du modèle.
– Burnham et Anderson [2004] recommandent d'utiliser le AICc quand $\frac{n}{K} < 40$.

5.2.3 Le critère d'Information de Schwarz (BIC)

Schwarz [1978] dérive le critère d'information bayésien :

$$\mathrm{BIC}_j = -2\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + K_j \log(n) \tag{5.13}$$

où BIC_j est le BIC du modèle $\mathfrak{M}_j$, $\mathcal{L}_j$ est sa fonction de vraisemblance, $\hat{\theta}_j$ est l'estimation de θ_j par maximum de vraisemblance, x représente les données empiriques générées à partir d'une vraie loi f, K_j est la dimension de θ_j et n est la taille de l'échantillon.

Bien que le BIC soit dérivé par Schwarz dans un contexte bayésien, il peut être utilisé en dehors de ce contexte car il ne dépend pas de la loi a priori.

Raftery [1995] montre que la probabilité a posteriori du modèle $p(\mathfrak{M}_j \mid D)$ sachant les valeurs observées (données D) peut être exprimée en fonction du BIC :

$$p(\mathfrak{M}_j \mid D) \approx \frac{\exp(-\frac{1}{2}BIC_j)}{\sum_{i=1}^{J}\exp(-\frac{1}{2}BIC_i)} \tag{5.14}$$

où D représente les données et J est le nombre de modèles testés.

Burnham et Anderson [2004] et Wagenmakers et Farrell [2004] présentent une variation de la probabilité a posteriori et l'appellent poids de Schwarz car elle est analogue aux poids d'Akaike :

$$w_j(\mathrm{BIC}) = \frac{\exp(-\frac{1}{2}\triangle \mathrm{BIC}_j)}{\sum_{i=1}^{J}\exp(-\frac{1}{2}\triangle \mathrm{BIC}_i)} \tag{5.15}$$

où $\triangle\mathrm{BIC}_j = \mathrm{BIC}_j - \min \mathrm{BIC}$. $\triangle_j(\mathrm{AIC})$ peut s'interpréter comme la perte d'information quand on ajuste un modèle $\mathfrak{M}_j$ au lieu de $\mathfrak{M}_{min}$. $w_j(\mathrm{BIC})$ est interprété comme la probabilité du modèle $\mathfrak{M}_j$ d'être le meilleur modèle parmi l'ensemble des modèles testés.

Quelques remarques

- Le BIC n'impose pas l'appartenance du vrai modèle à la famille des modèles ajustés.
- Le premier terme de l'équation (5.13) est la log-vraisemblance estimée au point x : elle augmente avec la qualité de l'ajustement du modèle à la donnée.
- Le deuxième terme de l'équation (5.13) pénalise les modèles trop paramétrés.
- La pénalisation du BIC est plus forte que celle du AIC. BIC décourage fortement la sélection des modèles avec beaucoup de paramètres.
- Le BIC est plus facile de calculer que le TIC.
- Raftery [1995] montre le BIC comme une approximation de la log vraisemblance marginale d'un facteur de Bayes BF (le rapport entre deux vraisemblances marginales des modèles différents).

5.2.4 Longueur de description minimale (MDL)

Cette méthode est basée sur la théorie du codage algorithmique utilisant l'information statistique contenue dans les données et les paramètres. Ce critère est défini par Rissanen [1978] et Rissanen [1983] :

$$\mathrm{MLD}_j = -\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + \frac{1}{2}\log \mid H_j(\hat{\theta}_j) \mid \tag{5.16}$$

où MLD_j est la MDL du modèle $\mathfrak{M}_j$, $\mathcal{L}$ est la fonction de vraisemblance de $\mathfrak{M}_j$, $\hat{\theta}_j$ est l'estimation de θ_j par maximum de vraisemblance, x représente les données empiriques générées à partir d'une vraie loi f, et $H_j(\hat{\theta}_j)$ est la matrice Hessienne de $-\log(\mathcal{L}_j(\hat{\theta}_j \mid x))$; elle est couramment appelé matrice d'information de Fisher observée.
On choisi le modèle dont le MDL_j est la plus petite.

Quelques remarques
- La MDL n'impose pas l'appartenance du 'vrai modèle' à la famille des modèles ajustés.
- Il est valide pour les modèles emboîtés et non emboîtés.
- Le premier terme est la log-vraisemblance estimée au point x : elle augmente avec la qualité de l'ajustement du modèle à la donnée.
- Le deuxième terme, contenant la matrice Hessienne pénalise la complexité du modèle par le nombre des paramètres et la forme fonctionnelle des modèles.
- Au fur et à mesure que la taille de l'échantillon augmente, la sensibilité à la forme fonctionnelle du modèle diminue : $\mid H_k(\hat{\theta}_j) \mid \rightarrow an^k$ quand $n \rightarrow \infty$ où a est une constante [Myung, 2000; Raftery, 1993]. Dans ce cas, MLD est reduit à $\frac{1}{2}$ BIC

5.2.5 Critère d'Information Complexe (ICOMP)

Bozdogan [2000] a développé cette mesure de la complexité du modèle. Le ICOMP est défini comme une mesure d'entropie de la dépendance statistique entre les valeurs estimées des paramètres :

$$\mathrm{ICOMP}_j = -\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + \frac{K_j}{2}\log\left[\frac{\mathrm{tr}(\Omega_j(\hat{\theta}_j))}{K_j}\right] - \frac{1}{2}\log \mid \Omega_j(\hat{\theta}_j) \mid \tag{5.17}$$

où ICOMP_j est le ICOMP du modèle $\mathfrak{M}_j$, $\mathcal{L}_j$ est sa fonction de vraisemblance, $\hat{\theta}_j$ est l'estimateur de θ_j par maximum de vraisemblance, x représente les données empiriques générées à partir d'une vraie loi f, $\Omega_j(\hat{\theta}_j)$ est la matrice de covariance des valeurs estimées des paramètres de $\mathfrak{M}_j$, et K_j est la dimension de θ_j.
Le modèle avec le ICOMP_j le plus petit sera choisi.

Quelques remarques
- Le ICOMP impose l'appartenance du vrai modèle à la famille des modèles ajustés.
- Il est valide pour les modèles emboîtés et non emboîtés.
- Le premier terme est la log-vraisemblance estimée au point x : elle augmente avec la qualité de l'ajustement du modèle à la donnée.
- Le deuxième terme pénalise les modèles trop paramétrés.
- Le troisième terme de l'équation pénalise la profusion de la complexité (interdépendance o corrélations entre les estimations des paramètres et l'erreur aléatoire du modèle.
- La correction exacte du biais dépend de la loi du modèle.

5.2.6 Critère d'information généralisé (GIC)

Le critère d'information généralisé a été proposé par Konishi et Kitagawa [1996] à partir de la théorie des fonctions d'influence; il vise à évaluer les modèles construits par diverses procédures d'estimation quand la famille spécifiée des distributions de probabilité ne contient pas nécessairement la distribution générant les données. Ce critère est obtenu par l'estimation de l'information de Kullback-Leibler mais est différent des critères précédents, en ce que le paramètre θ ne nécessite pas d'être estimé par maximum de vraisemblance. Landri [2004] le présente de la façon suivante : Pour l'instant on considère que $\hat{\theta}_j$ est un estimateur fonctionnel, à connaître. $\hat{\theta}_j$ est défini par : $\hat{\theta}_j = T(\hat{F})$ avec T une fonctionnelle régulière à p dimensions convenablement définie. On suppose que la fonctionnelle, sous ces considérations, est consistante au sens de Fisher, c.a.d., $T(G_\theta) = \theta$ pour tout $\theta \in \Theta$, où G_θ est la fonction de distribution de la densité de probabilité spécifiée $g(x \mid \theta)$.

Le GIC peut s'écrire comme suit :

$$\mathrm{GIC}_j = -2\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + \frac{2}{n}\sum_{i=1}^{n}\mathrm{tr}\left[T^{(1)}(x_i;\hat{F})\frac{\partial \log g(x_i \mid \theta)}{\partial \theta_j} \mid \hat{\theta}_j\right] \tag{5.18}$$

où GIC_j est le GIC du modèle $\mathfrak{M}_j$, $\mathcal{L}$ est sa fonction de vraisemblance, $\hat{\theta}_j$ est la valeur estimée de θ_j par une procédure quelconque, x représente les données empiriques générées à partir d'une vraie loi f, g est le modèle approximé en termes de probabilité, n est la taille de l'échantillon, et $T^{(1)}(x_i;\hat{F}) = (T_1^{(1)}(x_i;\hat{F}),...,T_p^{(1)}(x_i;\hat{F}))'$ est la fonction d'influence empirique à p-dimensions définie par :

$$T_i^{(1)}(X_i;\hat{F}) = \lim_{\epsilon\to 0}\frac{\left[T_i(1-\epsilon)\hat{F} + \epsilon\nu(X_i) - T_i(\hat{F})\right]}{\epsilon} \tag{5.19}$$

avec $\nu(X_i)$ masse au point X_i. La fonction d'influence $T^{(1)}(x;F)$ et son estimateur $T^{(1)}(X_i;\hat{F})$ sont respectivement les dérivées de $T(F)$ et de $T(\hat{F})$ par rapport aux mesures de probabilité $\nu(x)$ et $\nu(X_i)$. On considèrera comme meilleur modèle celui ayant le GIC le plus petit.

GIC pour des petits échantillons Comme dans le cas de l'AIC, le GIC de l'equation (11) est basé sur des approximations asymptotiques et n'est valable que pour les échantillons suffisamment grands. Ainsi Konishi et Kitagawa [2003] ont développé une théorie générale permettant l'obtention d'un critère d'information pour l'évaluation et la sélection de modèles de petits échantillons. Ce critère est le suivant :

$$\text{SGIC}_j = -2\log(\mathcal{L}_j(\hat{\theta}_j \mid x)) + 2\left\{b_1(\hat{F}) + \frac{1}{n}[b_2(\hat{F}) - \triangle b_1(\hat{F})]\right\} \tag{5.20}$$

où SGIC_j est le SGIC du modèle $\mathfrak{M}_j$, $\mathcal{L}_j$ est la fonction de vraisemblance, $\hat{\theta}_j$ est la valeur estimée de θ_j par une procédure quelconque, x représente les données empiriques générées à partir d'une vraie loi f, n est la taille de l'échantillon,

$$b_1(\hat{F}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{K_j} T_h^{(1)}(x_i;\hat{F})\frac{\partial \log f(x_i \mid \hat{\theta}_j)}{\partial\theta_h} \tag{5.21}$$

$b_1(\hat{F})$ est l'estimateur de $b_1(F)$, terme de correction de premier ordre dans la correction du biais asymptotique et
$b_2(\hat{F})$ est l'estimateur de $b_2(F)$, terme de correction de deuxième ordre dans la correction du biais asymptotique(Landri, 2004).

Quelques propriétés
- Le GIC n'impose pas l'appartenance du vrai modèle à la famille des modèles ajustés.
- Il est valide pour les modèles emboîtés et non emboîtés.
- Le premier terme est la log-vraisemblance estimée au point x : elle augmente avec la qualité de l'ajustement du modèle à la donnée.
- Le deuxième terme pénalise les modèles trop paramétrés.
- Le paramètre θ_j peut être calculé par n'importe quelle procédure : maximum de vraisemblance, maximum de vraisemblance pénalisé, des procédures bayésiennes, etc.

5.3 D'autres critères de sélection

5.3.1 Racine Carrée de l'Erreur Quadratique Moyenne (RMSE)

L'écart quadratique moyen ou erreur quadratique moyenne (RMSE) est une mesure fréquemment utilisée pour quantifier la différence entre les valeurs inférées par un modèle ou par un estimateur et les valeurs effectivement observées. Friedman et al. (1995) l'utilisent pour la comparaison des modèles. Myung (2000) présente une modification qui pénalise les modèles trop paramétrés :

$$\text{RMSE}_j = \sqrt{\frac{\text{SSE}_j}{n-K_j}} \tag{5.22}$$

où RMSE_j est la RMSE du modèle $\mathfrak{M}_j$, SSE_j est la somme des carrés des résidus de $\mathfrak{M}_j$, n est la taille de l'échantillon ajusté et K_j est le nombre des paramètres de $\mathfrak{M}_j$. Il faut noter que $(n - K_j)$ représente les dégrés de liberté du modèle. Le modèle dont le RMSE est le plus petit est choisi.

Quelques remarques
- C'est le critère le plus simple.
- Le numérateur mesure la qualité de l'ajustement du modèle à la donnée.
- Le dénominateur pénalise les modèles trop paramétrés.
- Le nombre de paramètres est la seule dimension de complexité considéré par ce critère.

5.3.2 C_p de Mallows

Le C_p est défini par [Mallows, 2000] :

$$\text{C}_{\text{p}_j} = \frac{SSE_j}{\hat{\sigma}^2} - n + 2K_j \tag{5.23}$$

où SSE_j est la somme des carrés des résidus de $\mathfrak{M}_j$, $\hat{\sigma}^2$ est l'estimateur de σ, n est la taille de l'échantillon ajusté et K_j est le nombre des paramètres du modèle j.
Le C_p est défini comme l'estimateur de l'erreur de prédiction au carré J_p :

$$J_p = \frac{E\left\{\sum_{i=1}^{n}(\hat{x}_i - E(x_i))^2\right\}}{\sigma^2}$$

En se basant sur ce critère, le meilleur modèle est choisi comme celui ayant le J_p le plus petit. Car J_p est en pratique une somme des carrés des résidus normalisée par σ^2, on peut voir le C_p comme une variation de la RMSE de la section 5.3.1.

5.3.3 Validation croisée (CV)

L'idée sous-tendant la validation croisée [Myung, 2000] est de procéder à une sélection de modèles, en se basant sur leur capacité à capturer le comportement des observations futures 'non encore collectées'. Pour les inférer, la base de données disponible est divisée en deux échantillons : l'un de calibration et l'autre de validation. Le premier est utilisé pour estimer les paramètres du modèle. Ensuite, les valeurs estimées sont fixées dans le modèle et sa capacité d'ajustement au second échantillon de validation est mesurée. À partir de ce résultat le coefficient de CV est obtenu :

$$\text{CV}_j* = D_{j*}(\hat{\theta}_j) \tag{5.24}$$

où CV_j est la CV du modèle $\mathfrak{M}_j$, D_{*j} es la mesure de défaut d'ajustement (e.g. SSE, % Var, log-vraisemblances, etc.), $*j$ est le modèle qui s'ajuste mieux et $\hat{\theta}_j$ est la valeur estimée du paramètre θ_j.
- On peut choisir n'importe quel critère pour la mesure.
- On suppose que le meilleur modèle est le vrai modèle.

5.4 Critères de choix de modèle à utiliser dans notre cas d'étude

Les critères basés sur la divergence de Kullback-Leibler proviennent d'une formulation mathématique du meilleur modèle que les autres critères n'ont pas. Nous choisissons donc les premiers. Entre eux, l'AIC et l'BIC sont les plus communs en écologie [Johnson et Omland, 2004]. L'AIC suppose d'avoir 'le vrai modèle' parmi les modèles testés. Dans la pratique, la grande complexité de l'écosystème associée à l'incertitude sur les processus permettent difficilement de faire cette supposition. Par conséquent, le BIC semble le meilleur critère pour les modèles du mouvement. Cependant, plusieurs auteurs ont utilisé l'AIC pour le choix de modèle pour le mouvement [Edwards *et al.*, 2007; James et Plank, 2007; Màrell *et al.*, 2002; Schooley et Wiens, 2003]. Pour cette raison, nous calculerons l'BIC et l'AIC pour tous les modèles. Puisque les nombres

des paramètres n'est varie beaucoup parmi les modèles (ils ont 1 ou 2 paramètres), il est probable que les deux critères donnent des résultats similaires.

6 Résultats

Les résultats seront présentés dans cet ordre : estimation des paramètres, tests d'ajustement et critères de choix de modèle.

6.1 Données et paramètres estimés

Rappelons que l'on suppose que les longueurs des déplacements de chaque bateau appartiennent à une population différente des autres. La table 1 montre le nombre des éléments et les valeurs minimale et maximale des queues de distribution des longueurs de déplacements pour chaque bateau. Bien que la valeur minimale ne varie pas beaucoup parmi les bateaux, la valeur maximale et la taille des échantillons sont très différentes pour chaque bateau. Des résultats très différents pour chaque navire sont donc à attendre.

	Bateau 1		Bateau 2		Bateau 2	
	$Q2$	$Q3$	$Q2$	$Q3$	$Q2$	$Q3$
n	2630	1315	586	293	2176	1088
x_{min}	2.64	7.83	2.23	4.94	3.01	8.54
x_{max}	255.18		39.95		1073.47	5

TAB. 1: Nombre d'éléments, valeurs minimale et maximale pour chaque queue de distribution pour chaque bateau.

Bateau	a	b	Log		Exp	ExpT	Pui	PuiT
			μ	σ^2	λ	λ	μ	μ
1	2.62	255.18	2.25	0.90	0.06	0.07	1.78	1.66
1	7.80	255.18	3.00	0.57	0.04	0.05	2.06	1.91
2	2.22	39.95	1.60	0.27	0.18	0.29	2.25	2.06
2	4.94	39.95	2.01	0.13	0.12	0.32	3.39	3.29
3	3.01	1073.47	2.62	1.79	0.02	0.02	1.66	1.58
3	8.54	1073.47	3.60	1.58	0.01	0.01	1.69	1.55

TAB. 2: Résultats de l'estimation des paramètres des lois proposées (log-normale, exponentielle, exponentielle tronquée, loi puissance et loi puissance tronquée), borne inférieure de la queue de distribution et troncature supérieure (pour les lois tronquées) pour distribution observée.

La table 2 montre les valeurs des paramètres estimées par maximum de vraisemblance. Les μ des lois puissance et puissance tronquée diffèrent certainement de ceux de l'article de Bertrand *et al.* [2007] pour les mêmes bateaux, estimées par régression linéaire simple des points des histogrammes des fréquences de longueurs des déplacements

6.2 Tests d'ajustement

Les figures 1, 2 et 3 montrent les graphiques Q-Q des quantiles des queues de distribution des bateaux versus les quantiles des lois de distributions proposées dont les valeurs des paramètres ont été estimées par maximum de vraisemblance. Puisque les valeurs des quartiles des lois proposées ont été basés sur des générations aléatoires des valeurs suivants les lois ; on a calculé aussi leurs intervalles de confiance. Les lignes rouges représentent les limites de confiance 0.05 et 0.95, et la ligne bleue représente la médiane. Pour le bateau 1 (Figure 1), les graphiques Q-Q de la log-normale, l'exponentielle et celui de l'exponentielle tronquée sont les plus proches d'une ligne droite. Les premiers quartiles, cependant, ne reporte pas la ligne droite de la continuation de la queue. Cela peut être dû à que la fréquence des observations de petites valeurs n'est pas assez grande, ce qui peut être dû au choix du début de la queue. On voit qu'effectivement, lorsque la queue des observations commence au troisième quartile, le problème se redue et le graphique pour la loi exponentielle tronquée est plus proche d'une ligne droite. Par contre aucune des deux ne s'ajuste bien aux petites valeurs de longueur des déplacements. La figure 4 montre l'ajustement des lois proposées aux queues de distribution du bateau 1. La petite courbe de la loi exponentielle ne contribue pas à son ajustement aux données (surtout pour la figure en bas), et les lois exponentielle et exponentielle tronquée semblent comme des meilleures alternatives. On voit ici encore que la loi exponentielle tronquée (en jaune) s'ajuste mieux à la queue de distribution qui commence au troisième quartile. Dans le cas de la loi puissance, les plus grandes valeurs affectent davantage l'ajustement de la loi aux données (Figure 2), ce qu'on ne peut pas noter dans la Figure 4, où les lois puissance (tronquées et non tronquées) semblent les meilleures candidates pour les données des deux queues de distribution.

Pour le bateau 2 (Figure 2), le graphique Q-Q de la loi puissance tronquée montre une ligne plus ou moins droite (queues de distribution commençant à la médiane et par le troisième quartile), aussi les lois exponentielle et exponentielle tronquée, surtout pour la queue de distribution qui commence à la médiane. Les lois puissances représentent mieux le comportement des faibles valeurs mais mal les très grandes valeurs (pour les deux queues de distribution ; Figure 5). Dans Figure 5, la loi exponentielle semble mieux s'ajuster aux données de la queue de distribution commençant à la médiane qu'à celles commençant au troisième quartile. La valeur du paramètre λ (0.175 et 0.123, respectivement) pour chaque queue de distribution conditionne une loi exponentielle de décroissance très lente (beaucoup plus pour la queue de distribution commençant au troisième quartile) et caractérise mal le comportement de la variable (longueur de déplacements) aux faibles valeurs. La loi exponentielle tronquée s'ajuste mieux aux valeurs petites que l'exponentielle, même si cela est insuffisant (Figure 5). Et elle ne s'ajuste pas bien aux valeurs très grandes à cause de sa vitesse de décroissance. La loi log-normale, par contre, ne s'ajuste bien à aucune des deux extrémités de la distribution (Figure 2 et Figure 5). Les valeurs grandes ont en réalité une probabilité plus fort d'arriver que celles-ci données par les loi exponentielle tronquée, puissance, puissance tronquée et log-normale.

Pour le bateau 3, aucune loi ne montre une ligne droite dans les graphiques Q-Q (Figure 3), probablement à cause des valeurs très grandes ; le bateau 3 a les valeur de longueur des déplacements les plus grandes parmi les trois bateaux (Table 1). Figure 6 montre un meilleur ajustement des lois puissance et puissance tronquée aux données par rapport aux autres lois.

La table 3 montre les résultats des tests d'ajustement pour les lois proposées. Toutes les lois sont rejetées pour le bateau 1 par tous les tests. Pour le bateau 2, la loi exponentielle tronquée est la seule qui n'a pas été rejetée, avec une p valeur de 0.03 pour le test d'Anderson-Darling et de 0.04 pour le test de Cramér-von Mises, ce qui est une 'acceptation' très faible de la loi exponentielle tronquée. On a précédemment décrit l'exponentielle tronquée comme un compromis entre l'exponentielle (plus adapté aux grandes valeurs de la variable et moins adapté aux petites valeurs) et les lois puissances (plus adaptées aux petites valeurs de la variable et moins adapté aux grandes valeurs). Le test MC-LR a accepté la loi puissance et la loi puissance tronquée pour le bateau 3 (p valeurs de 0.92 et 0.13, respectivement). C'est le seul à ne pas rejeter les lois. Quelques valeurs des longueurs de déplacements du bateau 3 sont des valeurs très grandes (de l'ordre de 600, 800, 1000, Figure 3). Le test MC-LR, robuste aux valeurs aberrantes [Brys *et al.*, 2008], considère

Bateau	Queue	AD	CMV	MC-LR
1	$Q2$	–	–	–
1	$Q3$	–	–	–
2	$Q2$	ExpT (0.03)	ExpT (0.04)	–
2	$Q3$	–	–	–
3	$Q2$	–	–	–
3	$Q3$	–	–	Pui,PuiT (0.92), (0.13)

TAB. 3: Résultats des 3 tests d'ajustement utilisés : AD (Anderson-Darling), CVM (Cramér-von Mises) et MC-LR (Test de 4.4). Pour chaque queue de distribution des longueurs de mouvements des bateaux on montre la loi qui n'a pas été rejetée dans chaque test (p>0.01) et la valeur p correspondante. Le calcul de la p valeur est détaille dans la section 4.5.

alors ces valeurs comme aberrantes. Comme nous avons déjà discuté pour les cas des autres bateaux dans des paragraphes précédentes, lorsque les valeurs les plus grandes perdent du poids, les lois puissances et puissance tronquées sont les meilleures à s'ajuster aux données.

Bateau	Queue	Loi	AD	CMV	MC-LR
3	$Q2$	ExpT	21.3	15.1	4049.9
		PuiT	21.6	15.8	3974.5
3	$Q3$	ExpT	10.7	7.7	1253.8
		PuiT	12.2	8.8	1277.6

TAB. 4: Temps de calcul des 3 test d'ajustement utilisés pour les lois exponentielle tronquée et puissance tronquée pour les données du troisième bateaux.

Pour comparer l'efficacité pratique des, tests nous avons calculé le temps de calcul nécessaire à chacun (Table 4) pour les lois tronquées : exponentielle et puissance, pour les deux queues de distribution des observations du bateau 3, dont la taille de l'échantillon est le plus grand. Le test de Cramér-von Mises, qui donne des p valeurs similaires à ceux du test d'Anderson-Darling, est le plus efficace dans tous les cas. Par contre, le test MC-LR, qui donne des résultats incompatibles avec les autres tests, prend beaucoup plus de temps.

6.3 Choix de modèle

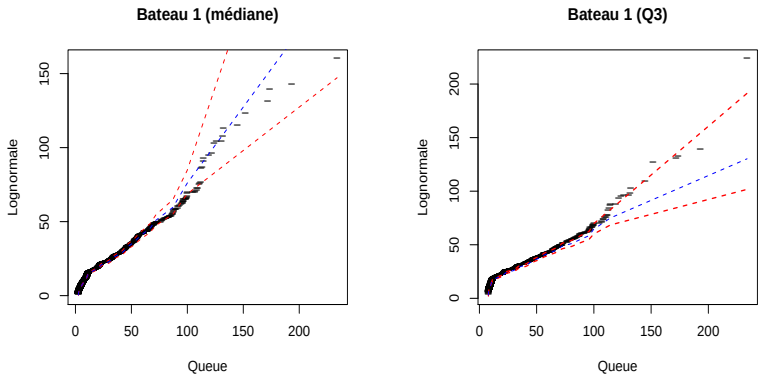
Comme déjà anticipé (section 5.4), les différences entre les valeurs des coefficients BIC et AIC sont très faibles et le choix du meilleur modèle est le même entre les deux méthodes dans tous les cas (Table 5 et Table 6). Nous notons que, pour les bateaux 1 et 2, le meilleur modèle est différent selon le point auquel on fait commencer la queue de distribution des observations : exponentielle tronquée lorsque la queue commence par la médiane et puissance tronquée lorsqu'elle commence par le troisième quartile. Pour le bateau 3, le choix est puissance tronquée dans les deux cas, une loi cependant rejetée par les tests d'ajustement.

Bateau	Queue	Log	Exp	ExpT	Pui	PuiT
1	$Q2$	19046.7 0.0	20749.3 0.0	**18957.3** **1.0**	19429.5 0.0	19839.3 0.0
1	$Q3$	11154.9 0.0	12103.9 0.0	10484.7 0.0	10390.0 0.0	**10305.1** **1.0**
2	$Q2$	3610.5 0.0	3675.5 0.0	**2642.2** **1.0**	2788.0 0.0	2747.0 0.0
2	$Q3$	2821.4 0.0	2174.4 0.0	1266.2 0.0	1261.9 0.1	**1257.5** **0.9**
3	$Q2$	19189.5 0.0	21494.7 0.0	20942.9 0.0	17608.4 0.0	**17494.2** **1.0**
3	$Q3$	11536.0 0.0	12201.2 0.0	11782.4 0.0	10840.8 0.0	**10732.2** **1.0**

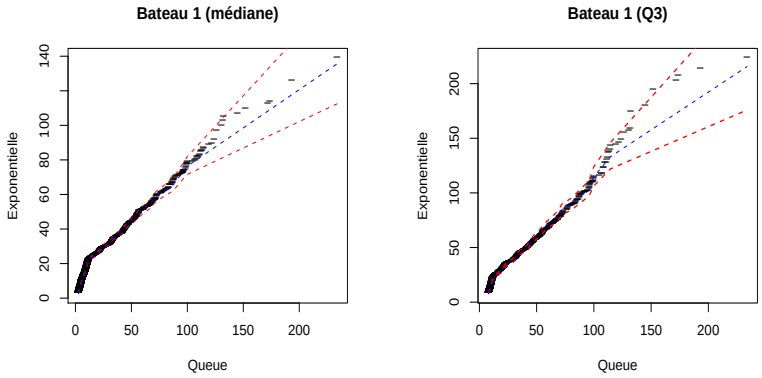
TAB. 5: Coefficient BIC et poids de Schwarz pour chaque loi utilisée pour chaque queue de distribution.

Bateau	Queue	Log	Exp	ExpT	Pui	PuiT
1	$Q2$	19035.0 0.0	20743.4 0.0	**18951.4** **1.0**	19423.6 0.0	19833.4 0.0
1	$Q3$	11114.6 0.0	12098.7 0.0	10479.5 0.0	10384.9 0.0	**10300.0** **1.0**
2	$Q2$	3601.8 0.0	3671.1 0.0	**2642.2** **1.0**	2783.6 0.0	2742.6 0.0
2	$Q3$	2814.0 0.0	2170.8 0.0	1262.6 0.0	1258.2 0.1	**1253.8** **0.9**
3	$Q2$	19178.2 0.0	21489.0 0.0	20937.2 0.0	17602.7 0.0	**17488.5** **1.0**
3	$Q3$	11526.0 0.0	12196.2 0.0	11777.5 0.0	10885.8 0.0	**10727.3** **1.0**

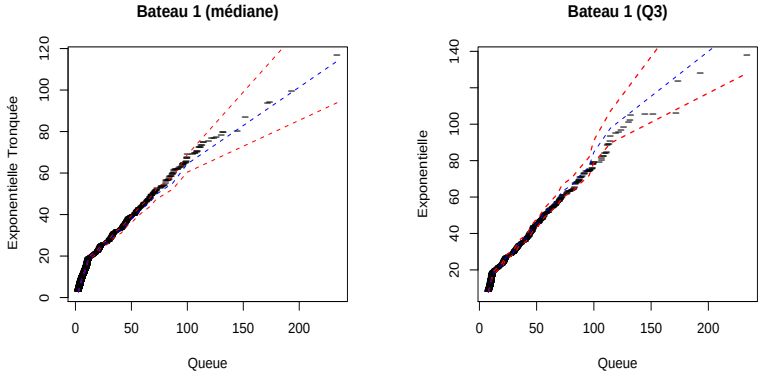
TAB. 6: Coefficient AIC et poids d'Akaike pour chaque loi utilisée pour chaque queue de distribution.


Bateau 1 (médiane)
Lognormale
Queue
0
50
100
150
200

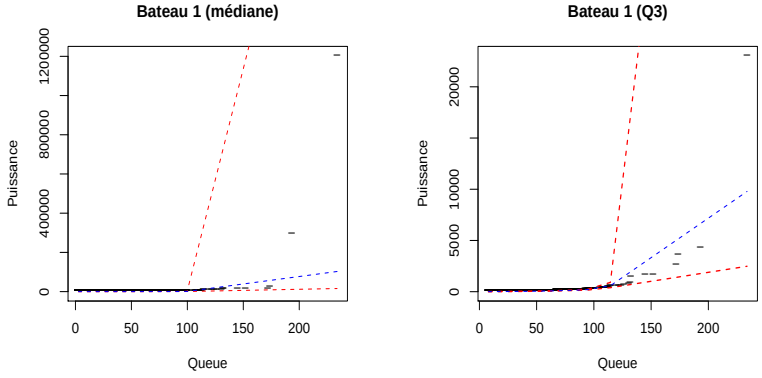
Bateau 1 (Q3)
Lognormale
Queue
0
50
100
150
200


Bateau 1 (médiane)
Exponentielle
Queue
0
20
40
60
80
100
140
0
50
100
150
200

Bateau 1 (Q3)
Exponentielle
Queue
0
50
100
150
200
0
50
100
150
200


Bateau 1 (médiane)
Exponentielle Tronquée
Queue
0
20
40
60
80
100
120
0
50
100
150
200

Bateau 1 (Q3)
Exponentielle
Queue
0
20
40
60
80
100
140
0
50
100
150
200


Bateau 1 (médiane)
Puissance
Queue
0
400000
800000
1200000
0
50
100
150
200

Bateau 1 (Q3)
Puissance
Queue
0
5000
10000
20000
0
50
100
150
200

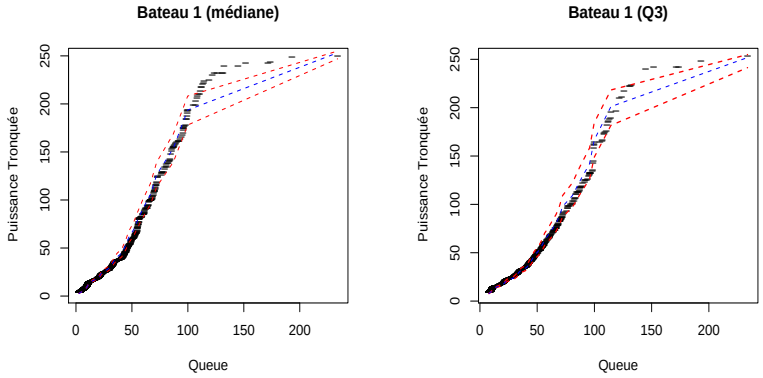
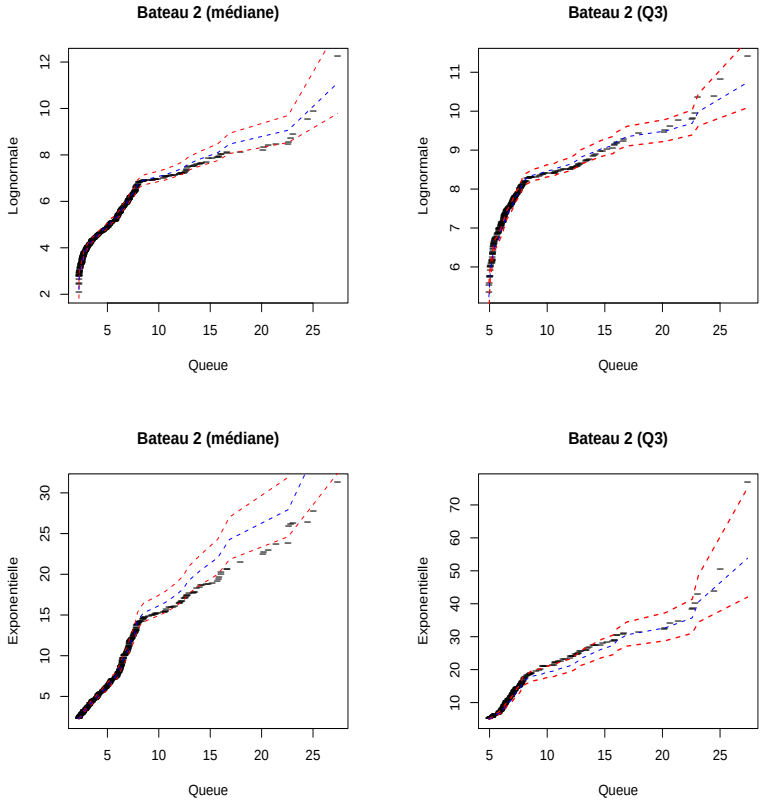


FIG. 1: Graphiques Q-Q pour les lois log-normale, exponentielle, exponentielle tronquée, puissance et puissance tronquée pour la queue de distribution des longueurs de mouvement du bateau 1. À gauche : queue de distribution à partir de la médiane. À droite : queue de distribution à partir du troisième quartile.



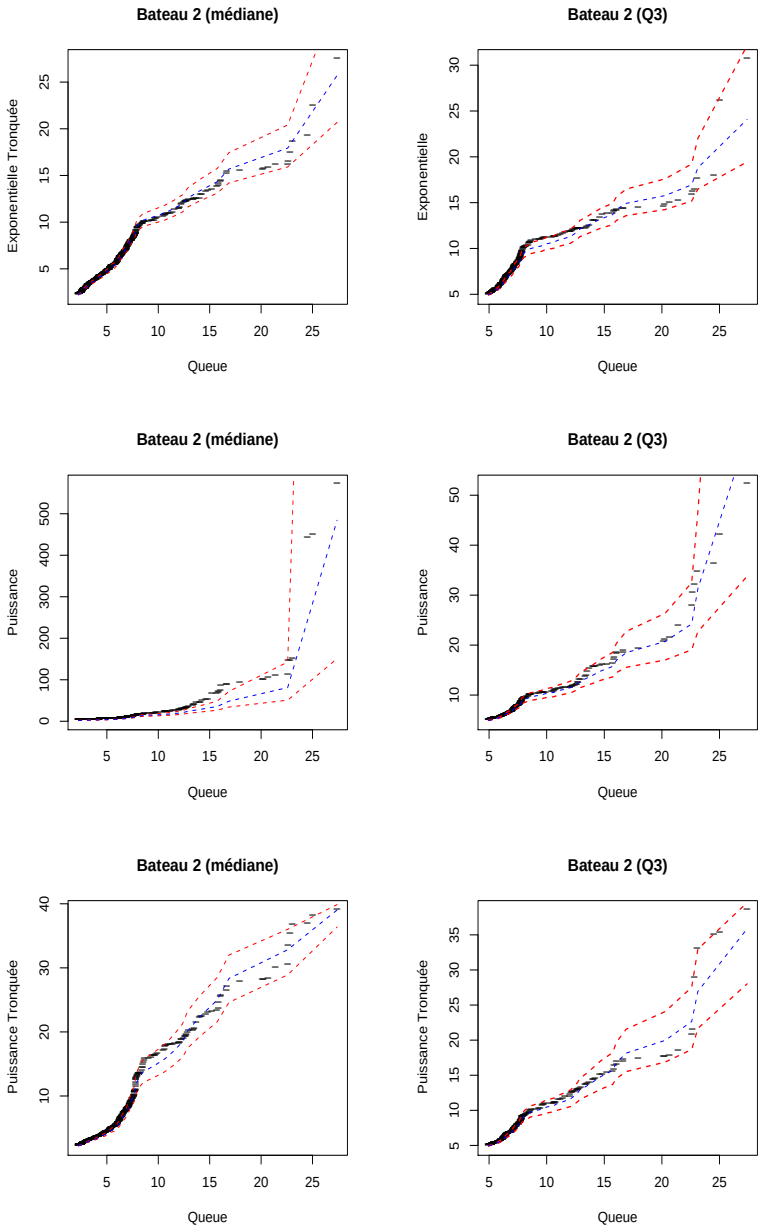
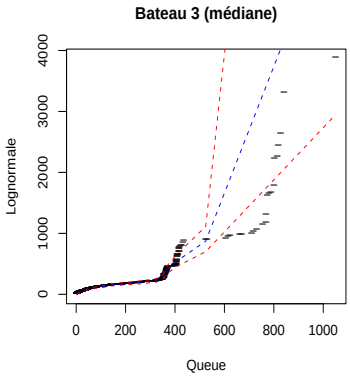
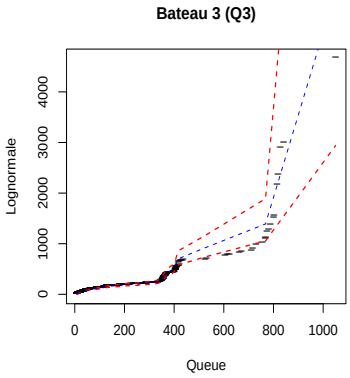
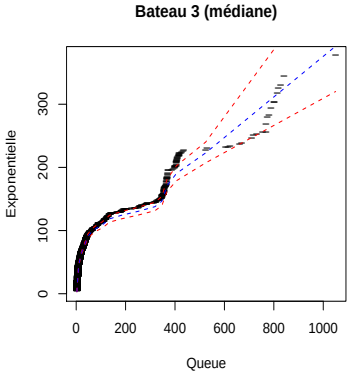
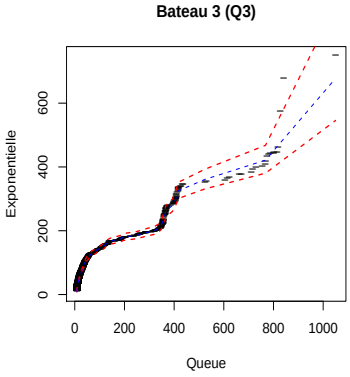


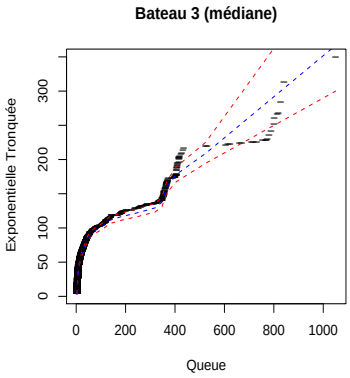
FIG. 2: Graphiques Q-Q pour les lois log-normale, exponentielle, exponentielle tronquée, puissance et puissance tronquée pour la queue de distribution du bateau 2. À gauche : queue de distribution à partir de la médiane. À droite : queue de distribution à partir du troisième quartile.

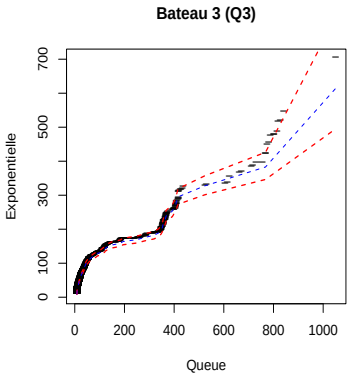

Bateau 3 (médiane)
Lognormale
0
1000
2000
3000
4000
0
200
400
600
800
1000
Queue


Bateau 3 (Q3)
Lognormale
0
1000
2000
3000
4000
0
200
400
600
800
1000
Queue


Bateau 3 (médiane)
Exponentielle
0
100
200
300
0
200
400
600
800
1000
Queue


Bateau 3 (Q3)
Exponentielle
0
200
400
600
0
200
400
600
800
1000
Queue


Bateau 3 (médiane)
Exponentielle Tronquée
0
50
100
200
300
0
200
400
600
800
1000
Queue


Bateau 3 (Q3)
Exponentielle
0
100
300
500
700
0
200
400
600
800
1000
Queue

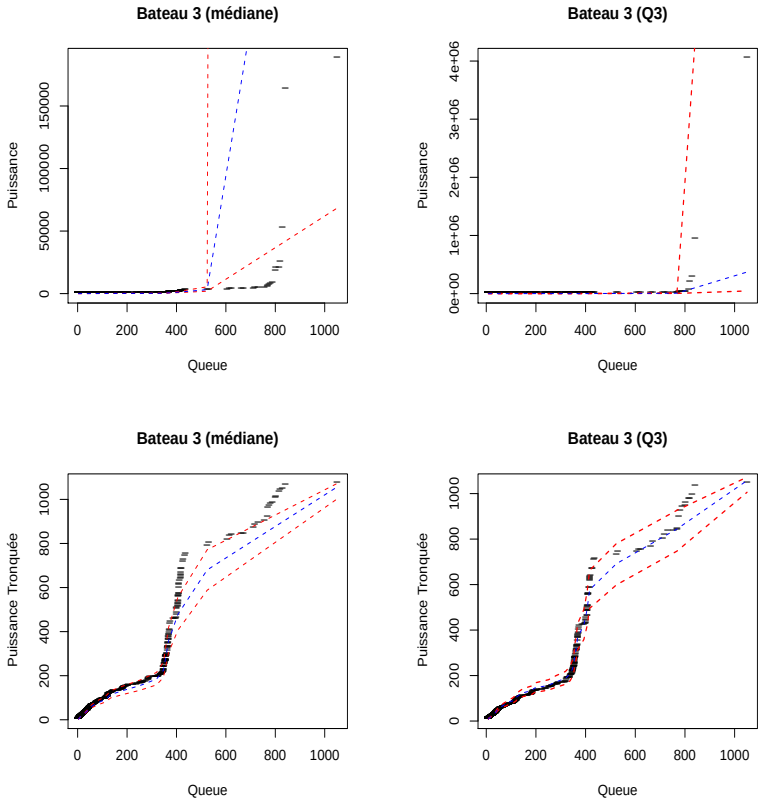


FIG. 3: Graphiques Q-Q pour les lois log-normale, exponentielle, exponentielle tronquée, puissance et puissance tronquée pour la queue de distribution du bateau 3. À gauche : queue de distribution à partir de la médiane. À droite : queue de distribution à partir du troisième quartile

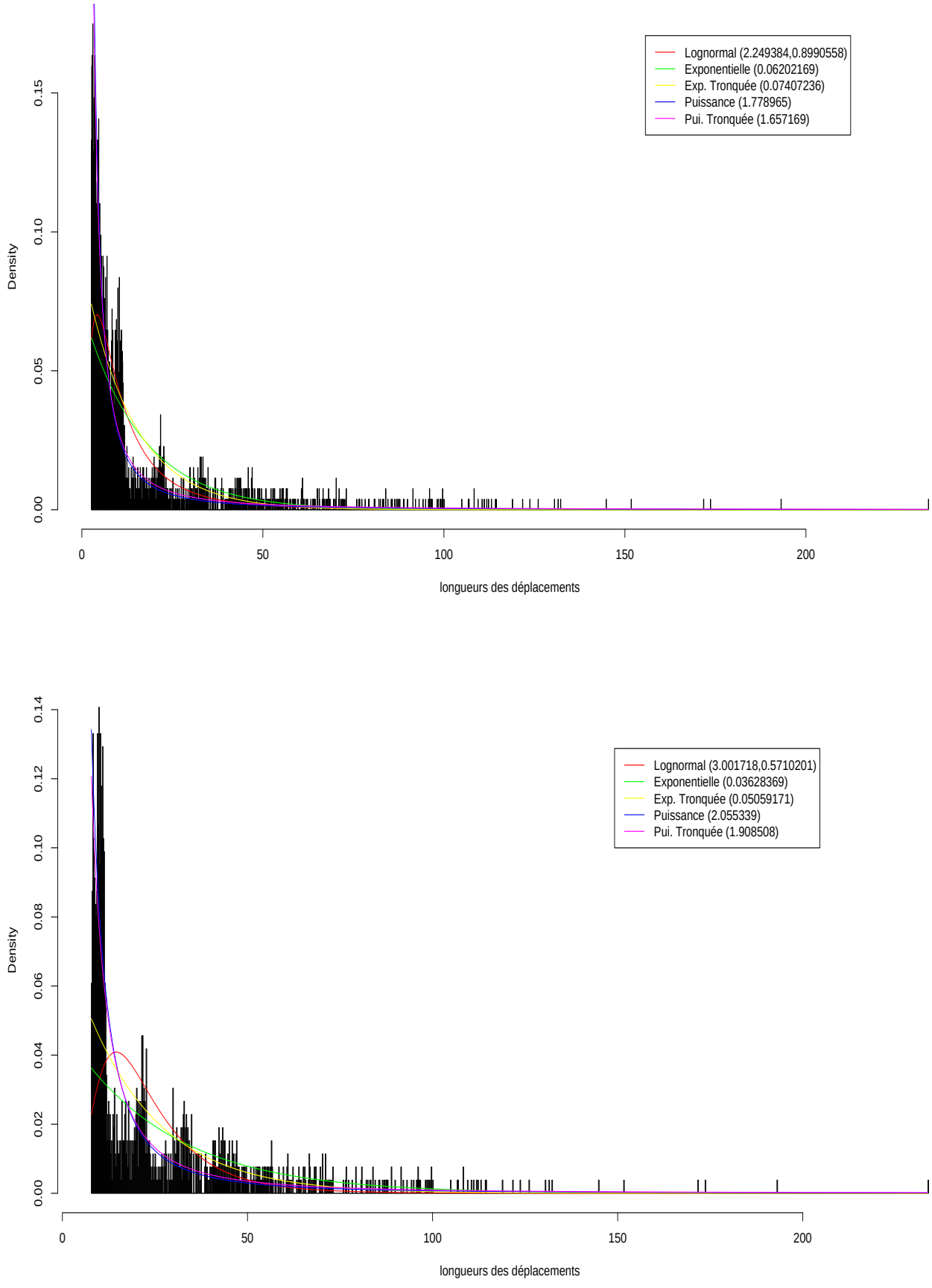


FIG. 4: Ajustement des lois à la queue de distribution du bateau 1. En haut : Queue de distribution commençant à partir de la médiane. En bas : Queue de distribution commençant à partir du troisième quartile.

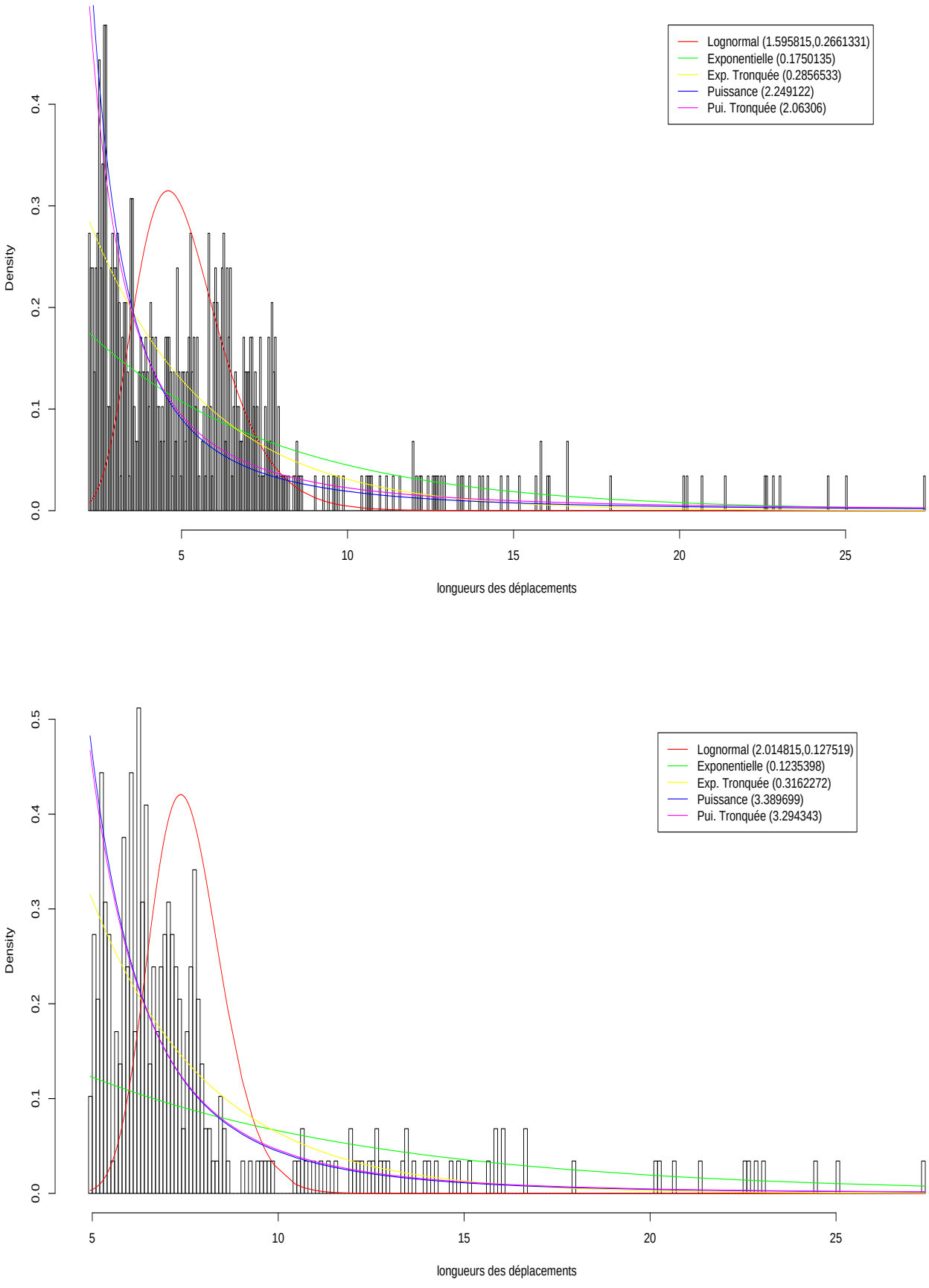


FIG. 5: Ajustement des lois à la queue de distribution du bateau 2. En haut : Queue de distribution commençant à partir de la médiane. En bas : Queue de distribution commençant à partir du troisième quartile.

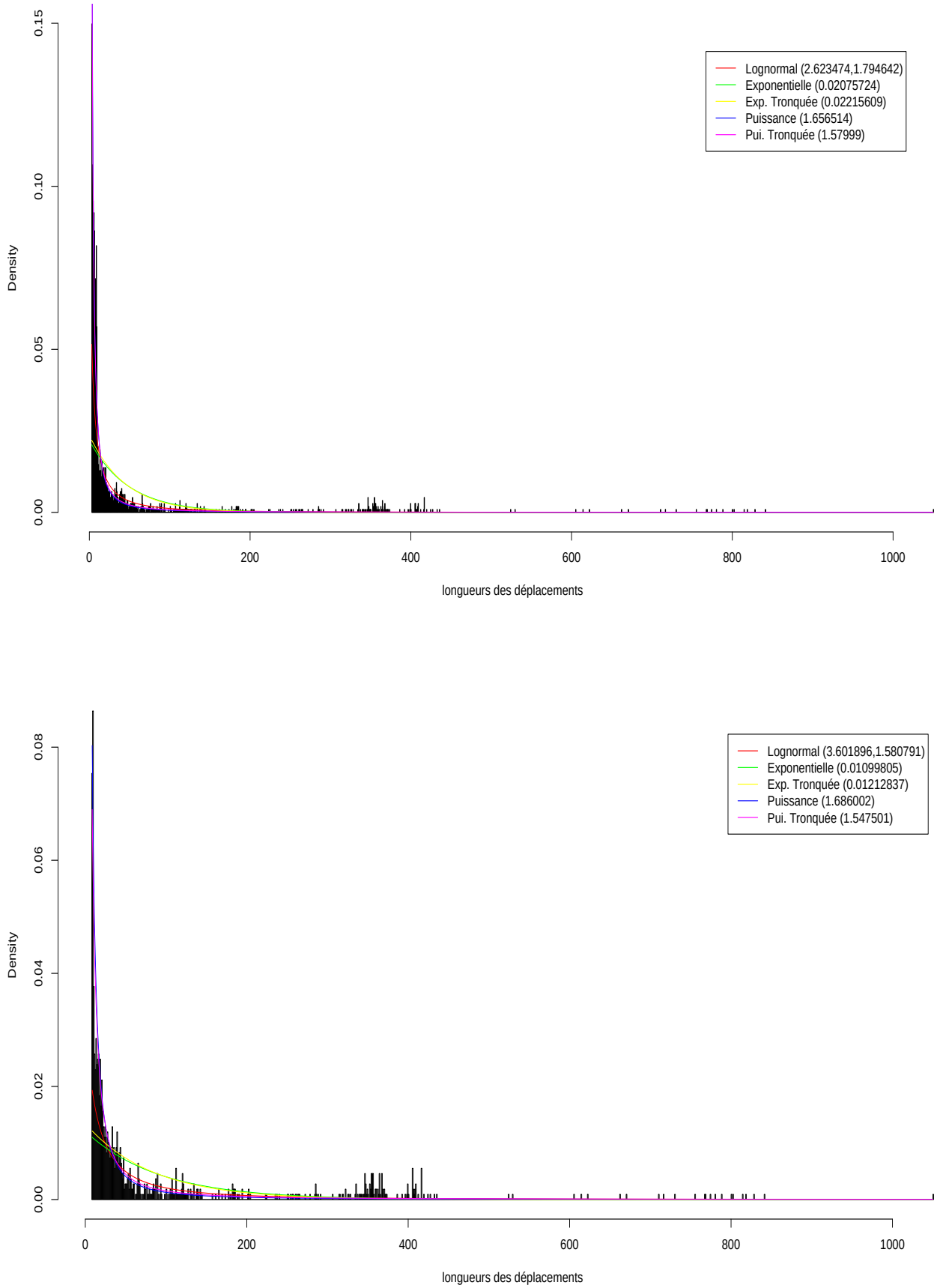


FIG. 6: Ajustement des lois à la queue de distribution du bateau 3. En haut : Queue de distribution commençant à partir de la médiane. En bas : Queue de distribution commençant à partir du troisième quartile.

7 Discussion et Conclusions

Dans ce chapitre nous discuterons les résultats obtenus dans le chapitre 6 : la modélisation des queues de distribution, la pertinence des tests d'ajustement pour des données réelles, le choix du critère de sélection de modèle et les implications des résultats en écologie.

7.1 Modélisation de la queue de distribution

Dans le chapitre 3 on a montré que les queues de distribution des longueur des déplacements des marches aléatoires de Lévy (mouvement Brownien y compris) sont des lois puissance (voir équation (3.8)). La diffusion et le type de processus de Lévy peuvent être caractérisés par la valeur du paramètre de la loi puissance (équation (3.1.1.2)), appelé indice de stabilité de Lévy. À partir de cette propriété, l'hypothèse de modélisation des longueurs des déplacements est approchée par l'hypothèse de modélisation de la queue des longueurs des déplacements par une loi puissance [Edwards, 2008]. Une loi puissance n'est pas définie en 0, une borne inférieur est toujours nécessaire. Dans ce travail nous avons utilisé d'autres lois alternatifs à la loi puissance pour la queue de distribution (borne inférieure, en tous cas) : loi puissance tronquée, exponentielle, exponentielle tronquée et log-normale. Les lois puissance tronquée et l'exponentielle tronquée répondent au fait que les longueurs de déplacements dans le monde réel ne sont pas infinies. Chaque individu (bateau) ne dépassera pas une longueur de déplacement maximum, déterminée par ses capacités motrices [Edwards, 2008].

Aucun critère théorique n'existe pour le choix du début de la queue de distribution, ce qui, en tous cas, n'a pas été considéré très important pour la modélisation jusque là [Edwards, 2008]. Dans notre application, on a choisi deux possibilités pour le début de la queue de distribution à modéliser : la médiane et le troisième quartile. Un choix plus restrictif, à partir du troisième quartile risquait de nous laisser très peu des données pour en faire de la modélisation (Figure 3). Les tables 5.10 et 5.13 montrent des différents choix de modèle quand la queue de distribution considérée change, ce qui est important car il montre la sensibilité de la modélisation aux choix du point de commencement de la queue de distribution.

7.2 Tests d'ajustement et de choix de modèle

7.2.1 Rejets par les tests d'ajustement

Les tests d'ajustement rejettent presque toutes les lois proposées. Pour les tests d'Anderson-Darling et Cramér-von Mises, la loi exponentielle tronquée, qui n'est pas rejetée pour la queue du bateau 2 qui commence à la médiane, a des p valeurs très faibles : 0.03 (AD) et 0.04 (CVM). Par contre, les lois non rejetées par le test de MC-LR, ceux-ci des lois puissance et puissance tronquée, ont des p valeurs plus grandes : 0.93 et 0.13. La faiblesse de la plus part des valeurs de p (des lois rejetées et non rejetées) n'est pas si grave : lorsque les données sont des données réelles et non simulées, on a une tendance à rejeter le modèle [G., 1990]. En plus, nous modélisons la queue de distribution des observations sans savoir vraiment où est-ce qu'elle commence. On est probablement très loin d'un 'vrai' modèle. Ce qui est plus relevant ici est alors, le choix de modèle. Dans ce contexte, où nous n'avons pas aucune idée du 'vrai modèle', ou d'un 'quasi-vrai' modèle [Lebarbier et Mary-Huard, 2006], le modèle sélectionné reste le meilleur modèle qui explique le comportement de notre variable, longueur de déplacements, parmi ceux qu'ont été comparés. Comme Box et Draper [1987] ont affirmé : 'essentiellement, tous les modèles sont faux, mais certains sont utiles'.

7.2.2 AIC versus BIC

Nous avons présenté les résultats des critères AIC et BIC. Lebarbier et Mary-Huard [2006] expliquent les différences de fond entre les deux : AIC est un critère de nature prédictif et la consistance du BIC en fait un bon critère pour des modèles explicatifs, comme c'est le cas. Dans la pratique, ayant des modèles qui n'ont pas plus de deux paramètres, on a constaté que les deux AIC et BIC donnent les mêmes choix des modèles dans tous les cas.

7.2.3 Coefficients versus poids

Le meilleur modèle est celui qui minimise le 'coefficient' BIC (ou AIC). Les poids d'Akaike et de Schwarz sont des transformations des BIC (ou des AIC) de telle manière qu'on obtienne des nombres qu'on puisse traiter comme des probabilités (voir sections 5.2.2 et 5.2.3). Logiquement, on choisit le modèle dont le poids est le plus grand. Les deux (coefficient et poids) donnent le même choix du meilleur modèle. Pour l'illustrer, les coefficients et les poids ont été montrés dans Table 6 et Table 5.

7.3 Modèles 'choisis' et leur implications en écologie

La loi puissance tronquée a été sélectionnée pour les 'deux queues de distribution' du bateau 3 ($\mu = 2.62$ et $\mu = 3.6$, respectivement), et pour les queues qui commencent par le troisième quartile des bateaux 1 ($\mu = 3$) et 2 ($\mu = 2.01$). D'un autre côté, la loi exponentielle tronquée a été choisie pour les bateaux 1 et 2 quand la queue de distribution commence à la médiane. Cela confirme le fait que les lois tronquées représentent mieux le mouvement (notons-nous que pour presque tous les modèles 'gagnants' l'ajustement aux données n'est pas si mauvaise selon l'analyse exploratoire).

Pour les cas où la loi puissance tronquée est préférée par rapport à l'exponentielle tronquée, la décroissance de la densité de probabilité est plus rapide. Les probabilités d'avoir des longueurs de déplacements petites sont très grandes et pour les longueurs des déplacements très grands, elles sont très petites. En termes de mouvement des bateaux de pêche, cela veut dire que les trajectoires sont composées par plusieurs déplacements courts et très peu des déplacements longs, ce qui est associé a des grosses agrégations d'anchois dans lesquelles on reste plus longtemps et dont on ne visite que quelques unes [Bertrand *et al.*, 2005]. Par contre dans une modélisation par une loi exponentielle tronquée, comme pour les deux cas déjà mentionnés, la décroissance de la densité de probabilité est plus lente. *Idem est*, il y a plus de déplacements courts que de longs, mais plus des déplacements longs que dans le cas de la loi puissance. Cela veut dire que les bateaux visitent plus des agrégations d'anchois ; cela peut être dû a des agrégations plus petites et plus disperses d'anchois.

En résumé, nous avons trouvé que la modélisation des queues de distribution est très sensible aux choix du début de ce que l'on considère être la queue de distribution. Le non ajustement des modèles aux queues de distribution n'est pas considéré comme rédhibitoire pour des données réelles, tandis qu'on n'a pas réellement d'idée du 'vrai modèle'. Nous concentrons alors notre attention sur le choix des modèles. Les résultats confirment que les lois tronquées sont des bonnes alternatives pour représenter les longueurs de déplacements, qui ne dépassent pas un maximum fixé pour chaque bateau. Pour le bateau 3, le meilleur modèle est celui de la loi puissance tronquée, pour les deux queues de distribution. Pour les bateaux 1 et 2, quand la queue de distribution commence par la médiane le modèle choisi est celui de la loi exponentielle tronquée et pour la queue de distribution qui commence par la médiane, celui de la loi puissance tronquée. Choisir une loi puissance tronquée veut dire que les trajectoires sont composées pour plusieurs déplacements courts et très peu des déplacements longs, indiquant qu'on reste longtemps dans les agrégations d'anchois et qu'on ne visite que quelques unes. Par contre, le choix d'une loi exponentielle veut dire qu'il y a plus des déplacements courts que longs, mais plus des déplacements longs que dans le cas de la loi puissance, indiquant qu'on visite plus des agrégations d'anchois où on ne reste pas beaucoup de temps.

Annexe : Rappels en Théorie de la Mesure

Définitions pris des notes des cours de Tran [2009] et de Campillo et Joannides [2009].

Definition .1 (Espace mesurable)
C'est un espace (ensemble) X que l'on munit d'une tribu.

Definition .2 (Tribu)
Une tribu ou σ -algèbre sur l'ensemble Ω est une famille $\mathcal{A}$ de sous-ensembles de Ω vérifiant

1. $\Omega \in \mathcal{A}$,
2. $\mathcal{A}$ est stable par passage au complémentaire, ca.d., $A \in \mathcal{A} \to A^c \in \mathcal{A}$,
3. $\mathcal{A}$ est stable par réunion dénombrable, c.a.d., si pour $n \geqslant 1$, $A_n \in \mathcal{A}$, alors $\bigcup_{n\geqslant 1} A_n \in \mathcal{A}$

Definition .3 (Mesure (positive))
Lorsque on a un espace mesurable, on peut associer à chaque ensemble A de la tribu $\mathcal{A}$ un poids appelé mesure de A et noté $\mu(\mathcal{A})$, tel qu'on ait :

1. $\mu(\Phi) = 0$,
2. Si (A_n) est une suite d'éléments deux à deux disjoints, alors $\mu(\bigcup_{n\geqslant 1} A_n) = \sum_{n\geqslant 1} \mu(A_n)$

Les probabilités sont des mesures telles que $\mu(X) = 1$. L'espace $(X, \mathcal{A}, \mu)$ est un espace mesuré.

Definition .4 (Mesure dominante)
On se place sur l'espace mesuré X muni d'une tribu $\mathcal{A}$. Si deux mesures μ et η sont telles que :

$$\forall A \in \mathcal{A}, \mu(A) = 0 \to \eta(A) = 0 \tag{.1}$$

alors μ est une mesure dominante de η.

Bibliographie

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, **16**, 3–14.

Applebaum, D. (2004). Lévy processes : From probability to finance and quantum groups. *Notices of the AMS*, **51**, 1336–1347.

Atkinson, R., Rhodes, C., Macdonald, D. W., et Anderson, R. M. (2002). Scale-free dynamics in the movement patterns of jackals. *Oikos*, **98**, 134–140.

Austin, D., Bowen, W., et McMillan, J. (2004). Intraspecific variation in movement patterns : modeling individual behaviour in a large marine predator. *Oikos*, **105**, 15–30.

Bakry, D. (2002). Modèles stochastiques. Cours de Maîtrise.

Bartumeus, F. (2007). Lévy processes in animal movement : An evolutionary hypothesis. *Fractals*, **15**, 151–162.

Bartumeus, F., Peters, F., Pueyo, S., Marrasé, C., et Catalan, J. (2003). Helical lévy walks : Adjusting searching statistics to resource availability in microzooplanckton. *PNAS*, **100**, 12771–12775.

Bartumeus, F., da Luz, M., Viswanathan, G., et Catalan, J. (2005). Animal search strategies : A quantitative random-walk analysis. *Ecology*, **86**, 3078–3087.

Beirlant, J., Goegebeur, Y., Teugels, J., et Segers, J. (2004). *Statistics of extremes : theory and applications*. Wiley, England.

Benhamou, S. (2007). How many animals really do the lévy walk ? *Ecology*, **88**, 1962–1969.

Bertrand, S., Burgos, J. M., Gerlotto, F., et Atiquipa, J. (2005). Lévy trajectories of peruvian purse-seiners as an indicator of the spatial distribution of anchovy (*Engraulis ringens*). *ICES Journal of Marine Science*, **62**, 477–482.

Bertrand, S., Bertrand, A., Guevara-Carrasco, R., et Gerlotto, F. (2007). Scale-invariant movements of fishermen : The same foraging strategy as natural predators. *Ecological Applications*, **17**, 331–337.

Box, G. et Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Boyer, D., Ramos-Fernández, G., Miramontes, O., Mateos, J., Cocho, G., Larralde, H., Ramos, H., et Rojas, F. (2006). Scale-free foraging by primates emerges from their interaction with a complex environment. *Proceeding of the Royal Society*, **273**, 1743–1750.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, **44**, 62–91.

Bradshaw, C., Sims, D., et Hays, G. (2007). Measurement error causes scale-dependent threshold erosion of biological signals in animal movement data. *Ecology Applications*, **17**, 628–638.

Brockmann, D., Hufnagel, L., et Geisel, T. (2006). The scaling laws of human travel. *Nature*, **439**, 462–465.

Brys, G., Hubert, M., et Struyf, A. (2004). A robustification of the jarque-bera test of normality.

Brys, G., Hubert, M., et Struyf, A. (2006). Robust measures of tail weight. *Computational Statistics and Data Analysis*, **50**, 733–759.

Brys, G., Hubert, M., et Struyf, A. (2008). Goodness-of-fit tests based on a robust measure of skewness. *Computational Statistics*, **23**, 429–442.

Burnham, K. et Anderson, D. (2004). Multimodel inference : Understanding aic and bic in model selection. *Sociological Methods and Research*, **33**, 261–304.

Bush, J., Woodruff, B., Moore, A., et Dunne, E. (1983). Modified cramer-von mises and anderson-darling tests for weibull distributions with unknown location and scale parameters. *Communications in Statistics - Theory and Methods*, **12**, 2465–2476.

Campillo, F. et Joannides, M. (2009). Processus stochastiques en temps continu pour la modélisation en écologie. Master2 Biostatistique.

Clauset, A., Rohilla Shalizi, C., et Newman, M. (2007). Power-law distributions in empirical data. cite arxiv :0706.1062.

Csörgó, S. et Faraway, J. (1996). The exact and asymptotic distributions of cramér-von mises statistics. *J. R. Statist. Soc. B*, **58**, 221–234.

Dubkov, A., Spagnolo, B., et Uchaikin, V. (2008). Lévy flight superdiffusion : An introduction. *Intern. Journ. of Bifurcation and Chaos*, **18**, 2649–2672.

Ducharme, G. (2009). Modèles paramétriques en biostatistique. Master2 Biostatistique.

Eberlein, E. (2001). *Application of generalized hyperbolic Lévy motions to finance.*

Edwards, A. (2008). Using likelihood to test for lévy flight search patterns and for general power-law distributions in nature. *Journal of Animal Ecology*, **77**, 1212–1222.

Edwards, A., Phillips, R., Watkins, N., Freeman, M., Murphy, E., Afanasyev, V., Buldyrev, S., da Luz, M., Raposo, E., Stanley, H., et Viswanathan, G. (2007). Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, **449**, 1044–1048.

G., S. (1990). *Probabilites, analyse des donnees et statistique. 2e Édition.* Editions Technip.

Galton, F. (1879). The geometric mean, in vital and social statistics.

Ghosal, S. et Roy, A. (2006). Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, **34**, 2413–2429.

Giles, D. (2000). A saddlepoint approximation to the distribution function of the anderson-darling test statistic. *Communications in Statistics - Simulation and Computation*, **30**, 899–905.

Goldstein, M., Morris, S., et Yen, G. (2004). Problems with fitting to the power-law distribution. *European Physical Journal B*, **41**, 255–258.

James, A. et Plank, M. (2007). On fitting power laws to ecological data. arXiv :0712.0613v1 [q-bio.QM].

Jarque, C. et Bera, A. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**, 255–259.

Johnson, J. et Omland, K. (2004). Model selection in ecology and evolution. *TRENDS in Ecology and Evolution*, **19**, 101–108.

Klafter, J., Zumofen, G., et Shlesinger, M. (1994). Lévy description of anomalous diffusion in dynamical systems.

Konishi, S. et Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Konishi, S. et Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection - functional approach. *Journal of Statistical Planning and Inference*, **114**, 45–61.

Kullback, S. et Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

Laherrere, J. et Sornette, D. (1998). Stretched exponential distributions in nature and economy : 'fat tails' with characteristic scales. *The European Physical Journal B. Condensed Matter and Complex Systems*, **2**, 525–539.

Landri, R. (2004). *Critères de Sélection de Modèles Non-Emboîtés*. Master's thesis, Université Montpellier II, Université Montpellier I et E.N.S.A. Montpellier.

Lebarbier, E. et Mary-Huard, T. (2006). Une introduction au critère bic : Fondements théoriques et interprétation. *Journal de la Société française de statistique*, **147**, 39–57.

Limpert, E., Stahel, W. A., et Abbt, M. (2001). Log-normal distributions across the sciences : Keys and clues. *BioScience*, **51**, 341–352.

Mallows, C. (2000). Some comments on cp. *Technometrics*, **20**, 87–94.

Mandelbrot, B. (1977). *The fractal geometry of nature*. Freeman, New York.

Mantegna, R. et Stanley, H. (1994). Stochastic process with ultraslow convergence to a gaussian : The truncated lévy flight. *Physical Review Letters*, **73**, 2946–2949.

Mårell, A., Ball, J., et Hofgaard, A. (2002). Foraging and movement paths of femlae reindeer : insights from a fractal analysis, correlated random walks, and lévy flights. *Canadian Journal of Zoology*, **80**, 854–865.

Massey, F. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, **46**, 68–78.

Metzler, R. et Klafter, J. (2002). From stretched exponential to inverse power-law : fractional dynamics, cole-cole relaxation processes, and beyond. *Journal of Non-Crystalline Solids*, **305**, 81–87.

Moors, J., Wagemakers, R., Coenen, V., Heuts, R., et Janssens, M. (1996). Characterizing systems of distributions by quantile measures. *Statistica Neerlandica*, **50**, 417–430.

Myung, J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190–204.

Newman, M. et Lutcavage, M. (2005). Power laws, pareto distributions and zipf's law. *Contemp.Phys*, **46**, 323–351.

Neyman, J. et Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference : Part i. *Biometrika*, **20**.

Nolan, J. (2010). *Stable Distributions - Models for Heavy Tailed Data*. Birkhäuser, Boston. In progress, Chapter 1 online at academic2.american.edu/~jpnolan.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Magazine, 5th Series*, **50**, 157–175.

Press, W., Teukolsky, S., Vetterling, W., et Flannery, B. (1992). *Numerical Recipes in C : The Art of Scientific Computing. 2nd edition*. Cambridge University Press, Cambridge, England.

Raftery, A. (1993). *Bayesian model selection in structural equation models.*

Raftery, A. (1995). *Bayesian Model Selection in Social Research.*

Ramos-Fernández, G., Mateos, J., Miramontes, O., Cocho, G., Larralde, H., et Ayala-Orozco, B. (2004). Lévy walk patterns in the foraging movements of spider monkeys (*Ateles geoffroyi*). *Behav. Ecol. Sociobiol.*, **55**, 223–230.

Rhee, I., Shin, M., Hong, S., Lee, H., et Chong, S. (2008). On the levy-walk nature of human mobility.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, **14**, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, **11**, 416–431.

Schervish, M. (1996). P values : What they are and what they are not. *The American Statistician*, **50**, 203–206.

Schooley, R. et Wiens, J. (2003). Finding habitat patches and directional connectivity. *Oikos*, **102**, 559–570.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Seier, E. (2002). Comparison of tests for univariate normality.

Shlesinger, M., West, B., et Klafter, J. (1987). Lévy dynamics of enhanced diffusion : Application to turbulence. *Physical Review Letters*, **11**, 1100–1103.

Shlesinger, M., Zaslavsky, G., et Klafter, J. (1993). Strange kinetics. *Nature*, **363**, 31–37.

Siegel, S. et Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences (2nd Ed.).* McGraw-Hill, New York, NY.

Sims, D., Righton, D., et Pitchford, J. (2007). Minimizing errors in identifying lévy flight behaviour of organisms. *Journal of Animal Ecology*, **76**, 222–229.

Sims, D., Southall, E., Humphries, N., Hays, G., Bradshaw, C., Pitchford, J., James, A., Ahmed, M., Brierley, A., Hindell, M., Morritt, D., Musy, M., Righton, D., Shepard, E., Wearmouth, V., Wilson, R., Witt, M., et Metcalfe, J. (2008). Scaling laws of marine predator search behaviour. *Nature*, **451**, 1098–1102.

Steele, M. (2002). *The power of categorical goodness-of-fit test statistics.* Ph.D. thesis.

Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, **7**, 13–26.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science*, **153**, 12–18.

Tran, V. (2009). Rappels de théorie de la mesure et de probabilité. École Nationale de la Statistique et de l'Administration Économique (ENSAE).

Travis, J. (2007). Do wandering albatrosses care about math? *Science*, **318**, 742–743.

Tsinober, A. (1994). Variability of anomalous transport exponents versus different physical situations in geophysical and laboratory turbulence.

Turchin, P. (1977). *Quantitative analysis of movement : measuring and modeling population redistribution in plants and animals.* Sinauer Associates, Sunderland, MA.

Varela, M., Ferraro, M., Jaroszewicz, S., et Mariani, M. C. (2005). Truncated levy walks applied to the study of the behavior of market indices.

Viswanathan, G., Afanasyev, V., Buldyrev, S., Murphy, E., Prince, P., et Stanley, H. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, **381**, 413–415.

Viswanathan, G., Buldyrev, S., Havlin, S., Da Luz, M., Raposo, E., et Stanley, H. (1999). Optimizing the success of random searches. *Nature*, **401**, 911–914.

Viswanathan, G., Afanasyev, V., Buldyrev, S., Havlin, S., da Luz, M., Raposo, E., et H.B., S. (2000). Lévy flights in random searches. *Physica A*, **282**, 1–12.

Viswanathan, G., Afanasyev, V., Buldyrev, S., Havlin, S., da Luz, M., Raposo, E., et H.B., S. (2001). Statistical physics of random searches. *Brazilian Journal of Physics*, **31**, 102–108.

Viswanathan, G., Bartumeus, F., Buldyrev, S., Catalan, J., Fulco, U., Havlin, S., da Luz, M., Lyra, M., Raposo, E., et Stanley, H. (2002). Lévy flight random searches in biological phenomena. *Physica A*, **314**, 208–213.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica*, **57**, 307–333.

Wagenmakers, E.-J. et Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic Bulletin Review*, **11**, 192–196.

Wilks, S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, **3**, 309–326.