



HAL
open science

Contribution to probabilistic and statistical modelling: stochastic models, statistical estimation, extremes and spatial models

Solym Manou-Abi

► To cite this version:

Solym Manou-Abi. Contribution to probabilistic and statistical modelling: stochastic models, statistical estimation, extremes and spatial models. Mathématiques [math]. Université de Montpellier, 2024. <tel-04906688v2>

HAL Id: tel-04906688

<https://hal.science/tel-04906688v2>

Submitted on 4 Sep 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC0 1.0 - Universal - International License

UNIVERSITE DE MONTPELLIER

Mémoire en vue d'obtention de
L'HABILITATION A DIRIGER DES RECHERCHES
Spécialité : Mathématiques et Applications

Contribution à la modélisation probabiliste et statistique : modèles
stochastiques, estimation statistique, extrêmes et modèles spatiaux

Présenté par Solym MANOU-ABI
le 05 juillet 2024

Devant le jury composé de

Jean-Noël Bacro
Sophie Dabo
Benoîte De Saporta
Toka Diagana
Adeline Leclercq Samson
Étienne Pardoux
Stephane Robin
Jean Vaillant
Anne-Françoise Yao

Professeur, Université de Montpellier
Professeur, Université de Lille
Professeure, Université de Montpellier
Professeur, University of Alabama in Huntsville
Professeure, Université de Grenoble
Professeur émérite, Aix-Marseille Université
Professeur, Sorbonne Université
Professeur, Université des Antilles
Professeure, Université Clermont Ferrand

Examineur
Examinatrice
Examinatrice
Invité
Rapporteuse
Examineur
Examineur
Rapporteur
Rapporteuse



UNIVERSITÉ
DE MONTPELLIER

Remerciements

Je remercie sincèrement Adeline Leclerq Samson, Jean Vaillant et Anne-Françoise Yao pour avoir accepté d'être rapporteurs pour cette Habilitation à diriger des recherches. Leur intérêt manifeste pour mon travail est pour moi un grand honneur, et je leur en suis très reconnaissant. Je remercie très sincèrement Stéphane Robin pour avoir accepté d'être examinateur à cette Habilitation.

Je tiens à exprimer ma gratitude envers Jean-Noël Bacro pour son précieux soutien scientifique, ainsi que pour son accompagnement dans mes projets de recherche. Je lui suis également reconnaissant pour ses conseils avisés sur la rédaction de ce manuscrit et pour avoir accepté d'être examinateur. C'est une chance pour moi de pouvoir bénéficier constamment de ton aide et soutien scientifique.

Malgré mon éloignement géographique par rapport à l'Institut Montpelliérain Alexander Grothendieck (IMAG), mon affiliation officielle pour la recherche à Montpellier, l'IMAG a joué un rôle essentiel dans mon développement scientifique. Je remercie ainsi tous les membres de l'IMAG qui m'ont offert des conditions agréables d'échange. Une mention spéciale à Benoîte De Saporta (qui m'a fait confiance dès mon arrivée à l'IMAG), de qui j'ai beaucoup appris et continue d'apprendre, tant sur le plan professionnel que sur le plan humain. Ses qualités scientifiques, son professionnalisme et ses qualités humaines sont une source constante d'inspiration. Merci pour ta participation au jury.

Je tiens également à exprimer ma reconnaissance envers l'ensemble de mes collègues de Mayotte, dont l'apport, qu'il soit direct ou indirect, a été précieux pour mon développement professionnel sur l'île. Je souhaite également adresser une reconnaissance à Christian Delhommé et Marianne Morillon du Laboratoire LIM, à l'île voisine de La Réunion, pour m'avoir intégré à leur équipe pédagogique du Master de mathématiques depuis quelques années.

Un très grand merci à Sophie Dabo d'avoir accepté de faire partie du jury et aussi pour sa très grande écoute en général. Ton impact scientifique a été d'une grande valeur pour ma carrière. Je te suis également reconnaissant pour nos échanges informels toujours enrichissants et appréciés. Merci à Étienne Pardoux, avec qui le contact a débuté au Sénégal et à Saint-Louis, pour m'avoir donné l'opportunité de recevoir une formation de qualité en probabilités à Toulouse. Je lui suis extrêmement reconnaissant pour le temps qu'il a consacré à participer à mon jury en qualité d'examineur.

Merci à Toka Diagana d'avoir accepté mon invitation de participer au jury ; c'est un honneur pour moi. Un grand merci pour l'accueil chaleureux que j'ai reçu lors de ma visite, il y a quelques années, à l'Université d'Alabama à Huntsville aux USA.

Un grand merci à Yousri Slaoui, avec qui j'ai débuté mes collaborations à Poitiers il y a quelques années. Ton envie et ton enthousiasme constants pour la recherche sont très enrichissants, tant du point de vue scientifique qu'humain.

Mon parcours est celui d'un autodidacte depuis mes travaux en Master recherche et la thèse à l'Institut de Mathématiques de Toulouse. J'en profite pour exprimer ma gratitude envers mes directeurs de thèse pour la rigueur scientifique qu'ils m'ont inculquée pendant cette période. Bien que je ne les aie pas revus depuis, leur impact sur ma formation perdure.

Vous tous, mentionnés ici, avez contribué à enrichir mes thématiques de recherche en mathématiques appliquées, que ce soit en épidémiologie, en santé-environnement, ou dans d'autres domaines. Cette interdisciplinarité entre probabilités, statistiques et modélisation est restée le fil conducteur de mes travaux de recherche, situés dans des thématiques variées. Cette interdisciplinarité convient non seulement à mes ambitions personnelles et professionnelles, mais surtout me permet de m'épanouir scientifiquement.

Parmi mes collaborateurs, Élodie, Benoîte, Sophie, Yousri, Angélo, El Hadji, Moustapha, Jean-Berky, Thomas et Bedreddine occupent une place toute particulière. J'ai énormément appris grâce à vous et pris plaisir à travailler avec vous. Merci pour toutes ces années de collaboration et d'échanges. Je remercie aussi tous mes autres collaborateurs pour les nombreux échanges passionnants et enrichissants. J'espère que nous continuerons encore longtemps à travailler ensemble.

Je remercie également tous les membres des laboratoires IMAG de Montpellier et LMA de Poitiers avec lesquels j'ai eu de nombreux échanges tant sur le plan humain et professionnel. Je remercie particulièrement André Mas, Jean-Michel Marin, Daniele A. Di Pietro, Alessandra Sarti et Boris Pasquier, qui m'ont offert des conditions de travail agréables.

Table des matières

1	Activités de recherche	7
1.1	Production scientifique et communications	8
1.2	Vie scientifique, Rayonnement et Responsabilités scientifiques	12
1.3	Responsabilités et participation dans des projets de recherche	15
1.4	Encadrement	16
2	Introduction générale	19
3	Equations différentielles stochastiques : existence et approximation	23
3.1	Existence de solutions	23
3.2	Approximation des solutions et processus stables	26
4	Statistique des processus stochastiques	33
4.1	Un modèle markovien pour étudier la dynamique de transmission de la fièvre typhoïde	33
4.2	Estimation des paramètres d'une solution d'EDS dirigée par un processus stable	40
5	Statistique computationnelle, statistique des extrêmes	45
5.1	Lois stables et estimation non paramétrique du taux de reproduction effectif	45
5.1.1	Estimation de paramètres de lois stables par fonctions caractéris- tique et score	46
5.1.2	Estimation de paramètres de lois stables par les algorithmes sto- chastiques	49
5.2	Estimation de lois à queues lourdes par la théorie des valeurs extrêmes .	53
6	Modélisation spatiale appliquée et apprentissage automatique	63
6.1	Modélisation spatiale de vecteurs d'arboviroses	63
6.2	Modélisation spatiale de données environnementales	67
7	Projets de recherche en cours	73
7.1	Modèles stochastiques : Estimation, Approximation et Apprentissage au- tomatique	73
7.2	Estimation des paramètres des lois stables par des statistiques des extrêmes	76
7.3	Estimation statistique, modèles spatiaux et Apprentissage Automatique avec des données réelles	79

Bibliographie**267**

Activités de recherche

Mes thématiques de recherche concernent à la fois les aspects théoriques des probabilités, la statistique mathématique et computationnelle incluant la modélisation en vue d'application à des données réelles. Les outils théoriques concernés sont les Chaînes de Markov, les lois à queues lourdes et processus stables, processus spatiaux, les équations différentielles stochastiques avec des approches d'existence de solutions, d'approximation et estimation paramétrique ou non paramétrique ainsi que la statistique des valeurs extrêmes. Je m'intéresse aux conditions de mélange, à l'inférence statistique, l'approximation numérique, à l'interpolation spatiale et à l'apprentissage automatique (Machine Learning). En termes d'applications, je m'intéresse aux données en santé-environnement (données de capture de vecteur d'arboviroses, de maladies infectieuses), épidémiologie (Covid-19, Dengue, Fièvre typhoïde, etc.), économétrie, finance quantitative (variation de prix).

1. **Analyse de Processus et Equations Differentielles stochastiques**

Equations différentielles, solutions périodiques et ses extensions, Equations différentielles stochastiques. Approximation numérique. Lois à queues lourdes, Chaînes de Markov, Processus de Lévy et processus α -stables.

2. **Estimation statistique : statistique computationnelle, statistique des processus, statistique des extrêmes**

Estimation de densités, Estimation paramétrique et non paramétrique pour des modèles stochastiques (EDS et Processus de Markov). Estimation des paramètres de lois à queues lourdes par la théorie des valeurs extrêmes.

3. **Modélisation spatiale et Apprentissage automatique (Machine Learning)**

Modélisation par des processus spatiaux en vue d'application en santé, environnement et finance.

Modélisation et apprentissage automatique en santé, écologie, environnement, économie, finance, etc.

Compétences informatiques

1. R studio, Python, Latex, Scilab, Rshiny
2. Analyse et visualisation de données avec R et Python

3. Statistique spatiale avec R et Python
4. Simulation avec R et Python
5. Cloud computing.

1.1 Production scientifique et communications

Je présente ici mes articles publiés, acceptés (à paraître), en révision et les articles soumis dans des revues à comité de lecture ; un chapitre d'ouvrage, un ouvrage collectif, des actes de conférences, articles (bulletins) de vulgarisation grand public.

Articles méthodologiques parus ou acceptés dans des revues internationales avec comité de lecture

- [1] Paul A. L. Faye, Elodie Brunel, Thomas Claverie, Solym M. Manou-Abi and Sophie Dabo-Niang. Automatic geomorphology mapping using statistical learning algorithms (2024). *Earth Science Informatics*, 1-18 (Springer).
- [2] Omar Hajjaji, Solym M. Manou-Abi and Yousri Slaoui. Parameter estimation for stable distributions and their mixture (2024). *Journal of Applied Statistics* 1–34 (Taylor & Francis).
- [3] Modou Kebe, El-Hadji Deme, Yousri Slaoui and Solym M. Manou-Abi. Robust estimator of the Ruin Probability in infinite time for heavy-tailed distributions. *Statistics*, 58(6), 1401-1422 (Taylor & Francis)
- [4] Solym M. Manou-Abi. Parameter estimation for a class of stable driven stochastic differential equation (2024). *Journal of the Indian Society for Probability and Statistics* (Springer). To appear
- [5] Mamadou Aliou Barry, E.H. Deme, Aba Diop and Solym M. Manou-Abi (2023). Improved Estimators of Tail Index and Extreme Quantiles under Dependence Serials. *Mathematical Methods of Statistics* (Springer) Vol. 32, 133–153.
- [6] Mamadou M. Mbaye and Solym M. Manou-Abi (2023). Existence of almost automorphic solution in distribution for a class of stochastic integro-differential equation driven by Lévy noise. *Journal of Analysis* (Springer) Vol. 31, 2139–2162.
- [7] Solym M. Manou-Abi and William Dimbour (2019). Asymptotically periodic solution of a stochastic differential equation. *Bull. Malays. Math. Sci. Soc* (Springer), 1-29.
- [8] Diouf M. B., Deme E. H., Manou-Abi M. S, Slaoui Y. (2024). Box-Cox transformation on the estimation of the extreme value index(EVI) and high quantiles for heavy-tailed distributions under dependance serials. Revised. *ALEA Lat. Am. J. Probab. Math. Stat.*
- [9] Mame Birame Diouf, Hadji Deme, Solym M Manou-Abi, Yousri Slaoui (2023) Kernel estimator of extreme value index (EVI) and high quantiles for heavy-tailed distributions under dependence serials using the Box-Cox transformation. *Afr. Stat.* 18(4) : 3651-3695.
- [10] William Dimbour and Solym M. Manou-Abi (2018). Asymptotically periodic functions in the stepanov sense and its application for an advanced differential equation with piecewise constant argument in a Banach space. *Mediterr. J. Math.* 15 : 25 (Springer).
- [11] William Dimbour and Solym M. Manou-Abi (2018). S-asymptotically periodic solution for a nonlinear differential equation with piecewise constant argument via S-Asymptotically periodic functions in the Stepanov sense. *Journal of Nonlinear Systems and Applications*. Vol 7 (1), 14-20.

[12] A. Joulin and S.M. Manou-Abi (2015). A note on convex ordering for stable stochastic integrals. *Stochastics* (Taylor & Francis), 87(4) : 1-12, 2015.

[13] P. Cattiaux and S. M. Manou-Abi (2014). Limits theorems for some functionals with heavy tails of a discrete time Markov chain. *ESAIM Probability and Statistics (EDP Sciences)*, 18 : 468-482.

Ouvrage collectif, chapitre d’ouvrage

[1] Solym M. Manou-Abi, Sophie Dabo-Niang and Jean-Jacques Salone. (Eds.) (2020) *Mathematical Modeling of Random and Deterministic Phenomena. Wiley.*

[2] Solym M. Manou-Abi, William Dimbour and M. Moustapha Mbaye (2020). *Asymptotically periodic solution for a stochastic fractional integro-differential equation. Mathematical Modeling of Random and Deterministic Phenomena. Wiley, 113-138.*

[3] Solym M. Manou-Abi, Sophie Dabo, Lema Logamou Seknewna, Julien Balicchi, Ambdoul-Bar Idaroussi. Spatio-temporal modeling and machine learning of mosquito abundance on the island of Mayotte. Preprint (2024). Under Revision.

Articles en révision dans des revues internationales avec comité de lecture

[1] Solym M. Manou-Abi. Approximate solution for a class of stochastic differential equation driven by stable processes. *Under revision.*

[2] Solym M. Manou-Abi, Emmanuel Gnandi, Said Said Hachim, Sophie Dabo-Niang, Jean-Berky Nguala. Comparative clustering methods : KL divergence, Rao distance, Bregman divergence with application to the Fani Maoré marine volcano earthquake data. Preprint 2024 *Under revision.*

Articles interdisciplinaires ou autre discipline parus dans des revues internationales avec comité de lecture

[1] William Dimbour and Solym M. Manou-Abi (2017). Asymptotically periodic solution for an evolution differential equation via periodic limit functions. *Int. J. Pure Appl. Math.* Vol. 113. 59-71.

[2] Solym M. Manou-Abi, Yousri Slaoui, and Julien Balicchi (2022). Estimation of some epidemiological parameters with the Covid-19 data of Mayotte. *Frontiers in Applied Mathematics and Statistics.* Vol. 8(67).

[3] Modou Kebe, El Hadji Deme, Abozou K. Tchilabalo, Solym M. Manou-Abi and Ebrima Sisawo (2023). Kernel estimation of the Quintile Share Ratio index of inequality for heavy-tailed income distributions. *European Journal of Pure and Applied Mathematics* Vol. 16(4), 2509–2543.

Proceedings

- [1] Solym Manou-Abi. P66-Estimation of some epidemiological parameters using the COVID-19 data of Mayotte. *Revue d'Épidémiologie et de Santé Publique* (Elsevier). Vol. 72, 2024
- [2] Ibrahim Bouzalmat, Benoîte de Saporta and Solym Manou-Abi(2023). Modélisation et estimation des paramètres d'un modèle multi-chaîne cachée de la fièvre typhoïde à Mayotte. *54es Journées de Statistique de la SFDS, la Société Française de Statistique (SFdS), Jul 2023, Bruxelles, Belgique.*
- [3] Ibrahim Bouzalmat, Benoîte de Saporta and Solym Manou-Abi (2022). Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte. *53èmes Journées de Statistique.*
- [4] Tsilefa, S. F, Manou-Abi, S.M. and Raheiririna (2022). Numerical convergence of some stochastic compartmental models in epidemiology. *Proceedings of African Symposium on Research in Computer Science and Applied Mathematics (CARI).*
- [5] Solym Manou-Abi and William Dimbour (2017). S-asymptotically periodic solutions in the p -th mean for a Stochastic Evolution Equation driven by Q -Brownian motion. *Proceedings of the International Conference on Applied Mathematics (ICAM2017), Taza, Morocco.*

Bulletins de vulgarisation du CNRS

- [1] Jean-Berky Nguala, Solym M. Manou-Abi, Angelo Raheiririna and Yousri Slaoui (2023). Jeux et arts traditionnels, supports d'apprentissage des mathématiques. *Microscoop, Octobre 2023*
- [2] Solym M. Manou-Abi, Yousri Slaoui and Julien Balicchi (2022). A Mayotte, une approche mathématique de la pandémie. *Microscoop, Mars 2022*

Preprints ou Articles soumis dans des revues internationales à comité de lecture

Les articles sont soumis dans des revues internationales bien connues avec comité de lecture affiliées à des éditeurs de renom.

- [1] Ibrahim Bouzalmat, Benoîte de Saporta and Solym Manou-Abi. Parameter estimation for a hidden linear birth and death process with immigration modeling disease propagation. Preprint (2023) *Soumis.*
- [2] Fridolin Melon, Solym M. Manou-Abi and Mahouton N. Hounkonnou. Parameter estimation for the $R(p, q)$ -trinomial probability distribution : properties and applications. Preprint (2024). *Soumis*
- [3] Solym M. Manou-Abi, Lema Logamou Seknewna, Sophie Dabo-Niang, Julien Balicchi, Ambdoul-Bar Idaroussi. Spatio-temporal modeling of mosquito abundance on the island

of Mayotte. Preprint (2023) *Soumis*

[4] Solym M. Manou-Abi, Essoham Ali, Yousri Slaoui and Julien Balicchi. Estimating social contact matrices from a Zero inflated Bell model through Density Power Divergence estimation : application to a sample survey data from the island of Mayotte. Preprint (2024)

[15] Guilherme Hilário Monteiro, Solym M. Manou-Abi, Bedreddine Ainseba and Stefanella Boatto. Modeling Weather-Driven Dynamics of Dengue Mosquitoes with Sparse Capture Data in Mayotte. Preprint (2024)

Liste des communications

Voici une sélection de quelques conférences internationales, nationales et séminaires.

Conférences internationales

- Conférence du RT MATRISK du 18 au 20 Juin 2024 au Mans (France). Stable driven stochastic models with applications in financial risk and epidemiology monitoring.
- Conférence internationale du 06 au 09 mars 2024. Statistiques et science des données, Université de Sfax, Tunisie. Estimation des paramètres pour des distributions stables à l'aide de statistiques de valeurs extrêmes.
- Conférence internationale du 13-17 décembre 2021 (conférencier invité). "Mathematical Modeling and Statistical Analysis of Infectious Disease Outbreaks", Dakar, Sénégal. *Titre de l'exposé : Modélisation de l'épidémie du Covid-19 à Mayotte par un modèle structuré en âge.*
- Conférence Internationale 14-16 décembre 2019. "12th International Conference of the ERCIM WG on Computational and Methodological Statistics, University of London", London, United Kingdom. *Titre de l'exposé : rate of convergence to α -stable laws in the generalized central limit theorem under the Zolotarev distance.*
- Conférence internationale 15-16 novembre 2018 "Colloque International sur la modélisation mathématique à Mayotte", Dembèni, Mayotte, France. *Titre de l'exposé : Solutions asymptotiquement périodiques d'équations différentielles stochastiques.*

Conférences nationales

- Conférence nationale, CORFEM du 9 juin 2023 (avec Jean-Berky Nguala) "XXIX èmes Journées de la Commission de Recherche sur la Formation des Enseignants de Mathématiques". Université de Nantes. *Atelier de recherche sur les jeux traditionnels Africains et leur portée mathématique.*
- Conférence nationale du 22 juin 2023 "Conférence ASIA 2023 Apprentissage Statistiques et intelligence artificielle". Université de Poitiers. *Titre : Modeling spatial dynamics of the Fani Maoré marine volcano earthquake data.*

- Conférence nationale du 27 Mai 2021 (en ligne). "Les pesticides et leur devenir dans l'eau, l'air et le sol (Congrès Français de recherche sur les Pesticides)" : *Titre : Analyse statistique et cartographique des déterminants pédologiques influençant la déchloration de la chlordécone en Martinique.*
- Conférence nationale du 11 décembre 2020. MASMSA 2020. Quatrième Rencontres Poitiers-Bordeaux : Algorithmes Stochastiques, Modélisation Statistiques et Applications", Poitiers, France. *Modélisation de l'épidémie du Covid-19 à Mayotte par un modèle structuré en âge.*
- Conférence du 19 Octobre 2018 au CUFR de Mayotte. Modélisation mathématiques : quels sont les objectifs et défis pour Mayotte ?

Séminaires

- Mercredi 11 octobre 2023 "Séminaire en ligne de Probabilités et statistiques du réseau AFRIMath", sur Zoom en visio-conférence. *Titre de l'exposé : Approximate solution for stable driven SDE.*
- Séminaire du 07 septembre 2022 "Séminaire en ligne de probabilités et statistiques du réseau AFRIMath", sur Zoom en visio-conférence. *Titre de l'exposé : Parameter estimation in a hidden birth and death process with immigration.*
- Séminaire du 15 Novembre 2022. Journées scientifiques de l'Institut de Recherche pour l'Enseignement des Mathématiques et des Sciences. *Titre de l'exposé : Modélisation du niveau et temps d'apprentissage des mathématiques à partir de variables socio-culturelles.* Centre Universitaire de Formation et de Recherche, Mayotte.
- Séminaire du 30 janvier 2020. Laboratoire de Mathématiques et Informatique, Université de La Réunion, Sainte-Marie, France. *Titre de l'exposé : L^p -bounded and stochastically continuous solutions of a class of semilinear stochastic evolutions equations driven by Lévy-stable processes.*
- Séminaire du 21 novembre 2019. Séminaire dynamique des populations, Institut de Mathématiques de Bordeaux, Bordeaux, France. *Titre de l'exposé : Intégrales et Equations Différentielles Stochastiques, presque automorphie et ordre convexe.*
- Séminaire du 15 Mai 2017. Séminaire de probabilités et statistiques, IMAG-Montpellier, Montpellier, France. *Titre de l'exposé : Autour des processus stables : théorèmes limites, ordres stochastiques et périodicité asymptotique.*

1.2 Vie scientifique, Rayonnement et Responsabilités scientifiques

Délégations

2023-2024 | Délégation CNRS d'une année à mi-temps au Laboratoire de Mathématiques et Applications. (LMA), Poitiers

2021-2022 | Délégation CNRS de 6 mois au Laboratoire de Mathématiques et Applications. (LMA), Poitiers

Séjours à l'étranger et collaborations

Mai 2021 Université Cheikh Anta Diop (Dakar) et Gaston Berger (Saint-Louis). Invitations pour collaborations de recherche avec Mamadou Moustapha Mbaye et El Hadji Deme, Enseignants-Chercheurs. (Senegal)

Mai 2019 Université de Fianarantsoa. Invitation à l'École Normale Supérieure de Fianarantsoa dans le cadre de collaborations pour recherche avec Angelo Raherinirina, Enseignant-Chercheur. (Madagascar)

Décembre 2019 Invité une semaine à l'Université d'Alabama in Huntsville (UAH) Collaboration pour recherche avec Toka Diagana, Professeur à l'UAH. (USA)

Novembre 2017 Invité deux semaines (18 au 30 novembre) à l'Université de Lomé (UL). Collaboration pour recherche avec Kossi Essona Gneyou, Professeur à l'UL. (Togo)

Expertise éditoriale

Pour des revues avec comité de lecture internationales en mathématiques et interdisciplinarité : évaluation et rapporteur d'articles de recherche.

- Fractional Calculus and Applied Analysis (Springer), Afrika Matematika (Springer), Plos One, Journal of Statistical Computation and Simulation (Taylor Francis).

Pour des conférences avec actes :

- Deux conférences CIMOM'2018 (Mayotte, Wiley) et ICAM'2017 (Maroc).

Représentation scientifique

Novembre 2023-présent | Responsable scientifique au Centre International de Mathématiques Pures et Appliquées. (CIMPA)

Octobre 2020-présent | Membre de l'IRN AFRIMath Réseau international de recherche du CNRS regroupant des mathématicien · ne · s localisé · e · s, principalement en Afrique subsaharienne et en France.

2016-2020 | Membre élu de la Commission de Recherche au CUFIR de Mayotte. évaluation et avis sur les projets de recherche soumis par les pairs.

Membre de comités de thèses

2020 | Rapporteur de la thèse d'Hamidou Ouedraogo. Université Nazi Boni, Burkina Faso
Titre : Modélisation mathématique et simulation numérique de systèmes dynamiques ordinaires et diffusifs en milieux marins.

Jury de concours de Recrutement

- 2023 : Président du Jury de recrutement des Educateurs Spécialisés. Rectorat de Mayotte, Juin 2023.
- 2023 : Membre du Comité de Sélection du poste PRAG-Mathématiques du Centre Universitaire de Mayotte, Mai 2023.
- 2020 : Membre du Comité de Sélection du poste MCF-26 Mathématiques et sciences du vivant du Centre Universitaire de Mayotte, Avril-Juin 2020.
- 2017 : Membre du Comité de Sélection du poste MCF 4212 de l'Université de la Réunion, site de Mayotte, Juin 2017.
- 2017 : Membre du Comité de Selection du poste PRAG-Mathématiques du Centre Universitaire de Mayotte, Juin 2017.
- 2017 : Membre du Comité de Sélection du poste PRAG-Informatique du Centre Universitaire de Mayotte, Novembre 2017.
- 2017-2019 : Présidence de jury au BAC à Mayotte.

Diffusion de la culture scientifique

Octobre 2023 | Animation stand. Fête de la Science, Poitiers (avec J-B. Nguala, MCF, Didactique des mathématiques). Laboratoire de Mathématiques et Applications.

Juin 2023 | Animation stand (avec Jean-Berky Nguala) d'un Atelier Sciences, Maths et Jeux traditionnels. Journées CORFEM à l'Université de Nantes.

2023-présent | Membre du comité de rédaction et de diffusion (grand public) du Panorama Santé. ARS de Mayotte.

Novembre 2022 | Animation stand lors de la Fête de la Science, Dembéné. CUFR et IREMIS de Mayotte.

2018-2019 | Journées Portes Ouvertes : présentation des offres de formation du CUFR au Forum des métiers. (Mayotte)

2017-2020 | Animation stand. Semaines de l'enseignement supérieur, Mayotte.

Organisation école de recherche, conférences, séminaires et implications

- 2024. Co-organisateur de l'école de recherche "Stochastic Modeling and Machine Learning" du 24 au 28 juin 2024 à l'Université Gaston Berger, Saint-Louis, Sénégal.
- 2023. Co-organisateur de la conférence ASIA 2023 "Apprentissage Statistiques et Intelligence Artificielle" du 20 au 23 juin 2023 au LMA, Université de Poitiers.
- 2021-2022. Co-organisateur, Responsable scientifique et administratif de l'Ecole de recherche CIMPA 2021 à Fianarantsoa, Madagascar (Novembre 2021 et Février 2022).
- 2021. Co-organisateur principal de deux journées Colloquium JMOD'21 : Premières Journées de Modélisation dans l'Océan Indien à Mayotte les 8 et 18 novembre 2021.

- 2021. Membre du comité scientifique et Conférencier invité à la conférence MASSE'21 : Méthodes Aléatoires pour les Sciences de la Santé, 11 au 15 décembre 2021 à l'Université Cheikh Anta Diop, Dakar, Sénégal.
- 2021-présent. Co-organisateur des séminaires en ligne de probabilités et statistiques et modélisation du réseau international IRN AFRIMath.
- 2019. Co-organisateur de l'école chercheur MIA (Mathématiques, Informatique et Applications) à l'Université de Fianarantsoa, Madagascar du 12 au 20 mai 2019.
- 2018. Organisateur principal du CIMOM'18 : Colloque International sur la Modélisation Mathématique à Mayotte du 15 au 17 Novembre 2018.
- 2018-présent. Organisateur des Séminaires de Mathématiques et Applications (SEMA) du Centre Universitaire de Mayotte.

1.3 Responsabilités et participation dans des projets de recherche

- 2021-présent : Participant de **MODCOV 19** plateforme de Modélisation et Covid-19 mis en place par l'INSMI et le CNRS <https://modcov19.math.cnrs.fr/pr>
- 2020-présent : Participant au réseau international de recherche du CNRS **AFRI-Math** (<https://www.afrimath.math.cnrs.fr/>) et réseau regroupant des mathématicien·ne·s localisé·e·s, principalement en Afrique subsaharienne et en France.
- 2020 : Co-porteur d'un projet de recherche intitulé **Modélisation probabiliste** et nommé **MODPROBA**. *Montant* : 49.000 euros, financé par la Commission de recherche du Centre Universitaire de Mayotte (45.000) et l'Agence Régionale de Santé de Mayotte (4000 euros). Ce projet a permis de co-financer un stage et la thèse de Ibrahim Bouzalmat. Thèse soutenue le 15 novembre 2023.
- 2021 : Porteur d'un projet de recherche **MODARS et STATFIR** (Analyse et Modélisation de données). *Montant* : 33.500 euros. Financé par l'Agence Régionale de Santé de Mayotte. Ce projet a engagé le recrutement de 4 stagiaires de Master2 et des chercheurs invités de l'Université de Poitiers, Lille et Montpellier.
- 2022 : Porteur d'un projet de recherche **MODSTAT** (Modélisation statistique). *Montant* : 50. 000 euros. Financé par l'Agence Régionale de Santé de Mayotte. Ce projet a engagé le recrutement d'un post-doctorant (Léma Logamou) et financé un CDD pour Omar Hajjaji.
- 2024 : Porteur d'un projet de recherche intitulé **MODRISQ** (Modélisation et caractérisation des risques sanitaires). *Montant* : 60. 000 euros. Financé par l'Agence Régionale de Santé de Mayotte. Ce projet est en cours de montage et d'étude.
- 2024 : Membre du projet de recherche portant sur **L'Etude de l'impact de la pluviométrie sur la propagation de la fièvre typhoïde à Mayotte**. Projet porté par Benoîte De Saporta (Université de Montpellier) et financé par l'Agence Régionale de Santé de Mayotte. *Montant* : 50. 000 euros.

Financement de la Recherche

J'ai participé au montage et à la coordination de projets de recherche pour lesquels des financements ont été obtenus auprès d'organismes institutionnels ou de recherche

comme : ARS de Mayotte , CIMPA, Union mathématique internationale (IMU), Conseil Départemental (CD) et Commission scientifique du CUFR de Mayotte (CS-CUFR).

Année	Organisme contractant	Montant (euros)	Nature du projet	Rôle
2019	ARS de Mayotte	4.000	Financement de stage	Porteur
2020	CS-CUFR	45.000	Co-financement-thèse	Porteur
2020	NUMEV-Montpellier	45.000	Co-financement-thèse	Associé
2020	CIMPA	12.000	Ecole de recherche	Porteur
2021	IMU	3.000	Ecole de recherche	Porteur
2021	CS-CUFR	3.500	Ecole de recherche	Porteur
2021	CD de Mayotte	3.500	Financement de stage	Porteur
2021-2024	ARS	35.500	Contrat MODARS	Porteur
2022-2024	ARS	50.000	Contrat MODSTAT	Porteur
2024-2027	ARS	60.000	Contrat MODRISQ	Porteur
2024-2027	ARS	50.000	Financement	Associé

1.4 Encadrement

Je présente ici les thèses que je co-encadre, suivi de post-doctorant effectué et une sélection de mémoires de stage encadrés.

Encadrement de thèses et postdoc

- 2022 – 2023 : **Léma Logamou Seknewna**. Contrat post-doctoral dans le cadre du projet de recherche MODSTAT du 15 septembre 2022 au 31 Août 2023. **Sujet** : Analyse et modélisation des données de santé à Mayotte. **Collaborateurs** : Jean-Noël Bacro et Sophie Dabo.
- 2020 – 2023 : **Ibrahim Bouzalmat**. Thèse de doctorat co-financé par le projet de recherche MODPROBA et le Labex NUMEV de Montpellier. **Thèse soutenue le 15 novembre 2023**. **Sujet** : Modélisation probabiliste de la dynamique de transmission de la fièvre typhoïde à Mayotte avec étude de risques épidémiques. **Directrice de thèse** : Benoîte De Saporta (Université de Montpellier). **Collaborateurs** : Julien Balicchi, ARS et CHU de Mayotte.
- 2020-présent : **Paul Aimé Faye**. Thèse de doctorat financé dans le cadre d'un contrat doctoral. Paul Aimé Faye est ATER à Lille pour l'année 2023-2024. **Sujet** : Méthodologie automatique de production de cartes géomorphologiques à l'aide d'algorithmes d'apprentissage statistique. **Directrice de thèse** : Elodie Brunel-Piccini. **Autres Co-encadrants** : Thomas Claverie (CUFR de Mayotte) et Sophie Dabo (Université de Lille).
- 2024-présent : **Stefana Tabera Tsilefa**. Thèse de doctorat à l'Université de Fianarantsoa. **Sujet** : Approximate almost periodic solutions in distributions and statistical estimation for stochastic models driven by stable processes in infinite space. **Co-encadrant** : Angelo Raheiririna (Université de Fianarantsoa) et William Dimbour (Université de Guyane).

- A partir de Septembre 2024 : **Gilles-Christ Dansou**. Lauréat d'une bourse de thèse Excellence France-Benin, prévue en cotutelle. **Sujet** : Recursive estimation of the drift and diffusion's terms for some stable-driven stochastic differential equations.

Encadrement : Post graduate student

Janvier-Juillet 2022. | Guilherme Hilario Monteiro (Université de Rio de Janeiro) Stage de recherche à Bordeaux. **Sujet** : Physiologically-Structured Population Dynamics for Dengue Mosquitoes in the island of Mayotte. En co-encadrement avec Bedreddine Ainseba (Bordeaux), Stefanella Boatto (Brésil) et Julien Balicchi (ARS-Mayotte)

Encadrement de stages de Master

J'ai encadré de nombreux étudiants en Master. Je présente ici quelques sélections.

- 2024 : **Sarah Ibrahim**, Stage de Master 2 du 02 février au 31 juillet. Université de Montpellier et Aix marseille. **Sujet** : Étude de l'impact de la pluviométrie sur la propagation de la fièvre typhoïde à Mayotte. **Co-encadrant** : Benoîte De Saporta (Professeure, Université de Montpellier).
- 2023 : **Abla Agbavito**, Stage de Master 2 de mars à août 2023 au CUFR de Mayotte. **Sujet** : Inférence statistique des paramètres des équations différentielles stochastiques.
- 2023 : **Gilles-Christ Dansou**, Stage de Master 2 de mai à juillet 2023 à l'IMAG de Montpellier. **Sujet** : Estimation non-paramétrique du terme drift dans des EDS dirigées par des processus α -stables.
- 2022 : **Omar Hajjaji**, Stage de Master 2 de mars à août 2022 au Laboratoire de Mathématiques et Applications, Poitiers. **Sujet** : Etude des modèles de mélange et lois stables. **Co-encadrant** : Yousri Slaoui (Université de Poitiers).
- 2022 : **Jeannot Thea**, Stage de Master 2 de mars à août 2022. Université de Lille. **Sujet** : Données spatiales, apprentissage statistique non paramétrique et applications. **Co-encadrants** : Sophie Dabo et Baba Thiam (Université de Lille).
- 2022 : **Abdallah Maghous**, Stage de Master 1, Université de Poitiers et CUFR de Mayotte. **Sujet** : Estimation statistique des paramètres des lois stables. **Co-encadrant** : Yousri Slaoui (Université de Poitiers).
- 2022 : **Floryan Renuy**, Stage de Master 2, Université de Lille et CUFR de Mayotte. **Sujet** : Analyse spatiale des données et cartographie des risques sanitaires à Mayotte. **Co-encadrants** : Sophie Dabo et Julien Balicchi.
- 2021 : **Kelly Moimba**, Stage de Master 1, Université de aLa Réunion. **Sujet** : Théorie de la ruine et processus de saut : modélisation, simulation et applications.
- 2020 : **Ibrahim Bouzalmat**, Stage de Master 2 de mars à août 2020 au CUFR de Mayotte et IMAG Montpellier. **Sujet** : Modélisation probabiliste de la dynamique de transmission des maladies hydriques. **Co-encadrants** : Benoîte De Saporta, Julien Balicchi (ARS de Mayotte).

- 2020 : **Paul Aimé Faye**, Stage de Master 2 de mars à août 2020 au CUFR de Mayotte. **Sujet** : Modélisation de la géomorphologie à l'aide d'algorithmes statistiques. **Co-encadrants** : Sophie Dabo et Thomas Claverie (Mayotte).

Introduction générale

Ce document présente une synthèse de mes travaux de recherche menés depuis ma thèse de doctorat au sein de l'Institut de Mathématiques de Toulouse et soutenue en Juin 2015. Mes travaux de recherche durant ma thèse tournaient autour des Chaînes de Markov et les propriétés de mélange en lien avec les lois et processus stables, ainsi que les inégalités de concentration convexe pour des intégrales stochastiques permettant de comparer des prix d'options. J'ai établi dans [77] des résultats de convergence de fonctionnelles markoviennes à queues lourdes sous des conditions d'indépendance asymptotique (conditions de mélange : fort mélange, rho-mélange et beta-mélange) vers des lois et processus à queues lourdes, notamment les lois et processus α -stables avec $\alpha \in (1, 2)$. Dans un second temps, dans [76], j'ai étudié les intégrales stochastiques dirigées par des processus α -stables. J'ai proposé des inégalités de comparaison convexe en adaptant un calcul stochastique forward-backward. Ensuite viennent de nouvelles orientations de mes travaux de recherche autour des données motivées par des problématiques en santé-environnement, épidémiologie et finance, etc, en terme de statistique des processus et statistique computationnelle, modélisation spatiale et la statistique des valeurs extrêmes pour des lois à queues lourdes. Ainsi mes travaux de recherche tournent autour des processus stochastiques, la statistique des processus et statistique computationnelle, la statistique des valeurs extrêmes et la modélisation. Ils contribuent à apporter d'autres résultats scientifiques dans le domaine des équations différentielles stochastiques, des chaînes de Markov, de l'estimation statistique, de la statistique des valeurs extrêmes et de la modélisation spatiale avec des applications sur données réelles. Ces travaux sont décrits dans ce document sous la forme de quatre chapitres.

Le chapitre 3, concerne l'existence et l'approximation numérique des solutions de modèles différentielles stochastiques incluant les processus stables. L'étude des solutions périodiques et ses extensions (asymptotiquement périodique, presque périodiques, automorphes, Bloch périodiques, etc) pour des classes d'équation d'évolution stochastiques en dimension infinie. Je me suis intéressé à diverses conditions d'existence et d'unicité de solutions presque automorphes, asymptotiquement périodiques et ses variantes dans le cadre de certaines familles d'équations (intégré) différentielles (stochastiques) dirigées par une famille large de processus stochastiques (de Wiener et Lévy). Par exemple, nous avons étudié dans [68] l'existence des solutions presque automorphes en loi pour des classes d'équations différentielles stochastiques dirigées par des processus de Lévy dans un espace de Hilbert. Ensuite dans [72] l'existence de solutions asymptotiquement périodiques pour des classes d'équations différentielles stochastiques dirigées par un mouvement dans un espace de Hilbert. Bien avant cela nous avons aussi établi dans [73, 74, 75] l'existence de

solutions asymptotiquement périodiques via les fonctions limites périodiques et asymptotiquement périodiques au sens de Stépanov pour une classe d'équations différentielles à argument constant par morceaux dans un espace de Banach. Deux projets de recherche sont actuellement en cours et concerne d'une part l'existence et l'approximation numérique de solutions d'une classe d'équations intégral-différentielles stochastiques dans un espace de Hilbert [100] et l'étude des solutions Bloch périodiques approximatives [101]. Les équations différentielles stochastiques correspondent souvent à des modèles mathématiques décrivant des phénomènes réels. Dans beaucoup de cas, on ne sait pas calculer une solution explicite et on doit utiliser des techniques de résolution approchée. On s'intéresse alors à la discrétisation et à la résolution numérique de l'équation. Des travaux de recherche ont été poursuivis dans le cadre de l'existence de solutions approximatives pour une classe d'équations différentielles stochastiques dirigées par des processus α -stables avec $\alpha \in (1, 2)$. Ainsi dans [86], j'étudie l'existence des solutions approximatives et fortes du schéma numérique d'Euler-Maruyama ainsi que les vitesses de convergence associées sous des conditions de type Lipchitz pour le terme de dérive et de type Holder pour le terme de diffusion. La méthode que nous proposons est basée sur une méthode de troncature en séparant les sauts du processus α -stable via la décomposition de Lévy-Itô. Nous donnons quelques simulations numériques de modèles stochastiques qui correspondent à nos résultats, à savoir les processus d'Ornstein-Uhlenbeck, de Cox-Ingersoll-Ross et de Lotka-Volterra perturbés par des processus α -stables.

Le chapitre 4 traite essentiellement de statistique de processus en vue d'applications sur données réelles en lien avec les processus stochastiques Markoviens et les solutions d'EDS dirigées par des processus stables. Concernant la partie d'estimation statistique et de modélisation markovienne, nous avons étudié le formalisme des processus de Markov pour la construction de modèles d'épidémie liées aux maladies hydriques notamment, la fièvre typhoïde. Nous avons proposé dans [93], un modèle de processus de naissance et de mort avec immigration pour modéliser la propagation la fièvre typhoïde et estimer les paramètres du processus lorsque la contamination provient à la fois du contact entre personnes et de l'environnement dans un cadre d'observations manquantes. D'autres travaux de modélisation markovienne ont été aussi menés également dans un cadre de collaboration avec la région océan indien de Mayotte notamment avec Madagascar. Ainsi que nous proposons dans [80] des modèles semi-markoviens pour la compréhension de la dynamique du Covid-19 à Madagascar basés sur des données réelles et simulées. Cet travail met en évidence un modèle semi-markovien prenant en compte les dimensions spatiales de la dynamique du Covid-19 dans les 22 régions de Madagascar. Concernant l'estimation statistique des solutions d'équations différentielles stochastiques dirigées par des processus stables (axes de recherche, j'ai développé dans [87], des méthodes d'estimation jointe des paramètres pour des solutions d'équations différentielles stochastiques dirigées par des processus α -stables ($\alpha \in (1, 2)$) et observées à des instants discrets à partir de l'estimateur de Nadaraya-Watson. En terme d'applications, nous considérons le problème de l'estimation conjointe des coefficients de dérive et de diffusion pour les processus de Cox-Ingersoll-Ross et d'Ornstein-Uhlenbeck dirigés par des processus α -stables avec $\alpha \in (1, 2)$.

Le chapitre 5 concerne l'inférence statistique de lois à densité, de lois stables et d'estimation statistique par la théorie statistique des valeurs extrêmes en vue d'applications aux données réelles. Dans un cadre de modélisation statistique de l'épidémie du Covid-19, un article de vulgarisation a été publié dans Microscop [79] (revue de vulgarisation scientifique du CNRS, délégation de Poitou Charentes) . D'autres travaux ont été poursuivis

dans l'article [71] où nous avons proposé des méthodes d'estimation statistique par la méthode de mélange du taux de reproduction effectif par une formulation non paramétrique existante. J'ai poursuivi cette activité de recherche par la prise en compte de la présence d'évènements avec probabilité faible (par exemple interval sériel négatif) en introduisant les lois α -stables dans [90] avec $\alpha \in (0, 2)$. Une perspective de recherche de ces travaux est actuellement en cours [97] en utilisant les méthodes d'estimation par la théorie des valeurs extrêmes. Dans la continuité des travaux d'estimation statistique par la théorie des valeurs extrêmes pour des lois de distributions à queues lourdes, nous avons traité le problème de l'estimation de l'indice des valeurs extrêmes basé sur la méthodologie du Jackknife généralisé dans [69]. Nous avons aussi proposé dans [70] des classes d'estimateurs semi-paramétriques de l'indice QSR (quintile Share Ratio) pour des distributions de revenus à queues lourdes. Nous nous sommes aussi intéressés aux probabilités de ruine d'une compagnie d'assurance pour des pertes d'assurance provenant de distributions à queues lourdes dans [85].

Le chapitre 6 traite de modélisation spatiale appliquée. Dans un cadre de modélisation spatiale en biologie marine, on s'intéresse à la mise en place de méthodes spatiales dans la production des cartes géomorphologiques marines, utiles pour la gestion des ressources ou de la planification de la conservation. Bien que les techniques de construction de ces cartes soient de plus en plus sophistiquées, les techniques manuelles sont encore largement utilisées. Des approches automatisées d'apprentissage statistique sont nécessaires pour obtenir des cartes reproductibles en un temps raisonnable. Un article [92] a été élaboré dans ce sens. Dans le cadre de la modélisation spatiale en santé-environnement, des projets de recherche ont été menés en collaboration avec l'Agence Régionale de Santé de Mayotte. Par exemple dans le cadre de la surveillance des arboviroses comme la dengue, nous avons introduit des ratios d'abondance de risque spatio-temporel dans [88] dans le but de fournir des probabilités d'abondance et les zones géographiques où les services de santé pourraient appliquer des mesures de surveillance et de contrôle des vecteurs. Ce travail fournit une modélisation statistique spatio-temporelle pour les données de comptage des moustiques *Aedes* en phase Oeuf et Adulte prenant en compte les conditions environnementales dans le contexte de l'épidémiologie des moustiques. Toujours dans un cadre de modélisation spatiale, j'ai également mené un travail de modélisation des données sismiques issues de la crise sismo-volcanique dans l'est de Mayotte de 2018 à 2021. Nous avons mis en évidence dans [84] des modèles basés sur les processus spatiaux et spatio-temporels ainsi que les séries temporelles. L'objectif étant de comprendre la dynamique de l'évolution spatiale de l'activité sismique à Mayotte.

Dans la suite de ce manuscrit, je propose une analyse d'une partie des travaux scientifiques mentionnés. J'ai fait le choix de présenter des travaux de recherche incluant au moins un axe des trois grandes thématiques de recherche mentionnées. Pour finir, je présente une partie de mon projet de recherche.

Equations différentielles stochastiques : existence et approximation

Je présente dans cette section les travaux menés sur les équations différentielles, équations différentielles stochastiques dirigées par des processus de Lévy y compris le mouvement Brownien (dans des espaces de Banach et Hilbert) et les processus α -stables. Les publications relatives à ces travaux se retrouvent dans [68, 72, 73, 74, 75, 81, 82]. Dans un premier temps, il s'agira de conditions d'existence de solutions asymptotiquement périodiques, presque automorphe en loi qui sont des variantes de la périodicité. J'ajouterai des résultats sur l'approximation numérique des solutions d'équations différentielles stochastiques dans le cadre où elles sont dirigées par des processus α -stables.

3.1 Existence de solutions

Les équations différentielles stochastiques (EDS) sont une généralisation des équations différentielles prenant en compte un terme de bruit blanc. Le concept de périodicité et de ses variants (presque-périodicité, périodicité asymptotique, etc) a été développé en relation avec des problèmes liés aux équations différentielles. Les fonctions presque périodiques formalisent des oscillations dont les fréquences ne peuvent se réduire à une seule fréquence de base. Les définitions des fonctions presque périodiques $f : \mathbb{R} \rightarrow \mathbb{X}$ avec \mathbb{X} un espace de Banach, font intervenir une norme $\|\cdot\|$ choisie sur l'espace des fonctions f et trois notions.

- Les presque-périodes : pour tout $\epsilon > 0$ il existe $l > 0$ tel que pour tout $a \in \mathbb{R}$, il existe un $\tau \in [a, a + l[$ de sorte que $\|f(\cdot + \tau) - f(\cdot)\| \leq \epsilon$.
- La propriété de Bochner : de toute suite de translatées $(f(\cdot + s_n))_n$, on peut extraire une sous-suite convergente pour la norme $\|\cdot\|$.
- La propriété d'approximation : les fonctions presque-périodiques au sens de Bohr sont les limites uniformes sur toute la droite réelle de polynômes trigonométriques de la forme $P_n(t) = \sum_{k=-n}^n a_k e^{i\lambda_k t}$ où l'ensemble des fréquences $\{\lambda_k; k \in \mathbb{Z}\}$ est une suite numérique quelconque.

Par exemple la fonction numérique f définie par $f(t) = \cos(t) + \cos(2\pi t)$ est presque-périodique sans être périodique (les périodes 2π et 1 sont en rapport irrationnel). A partir de 1928, Jean Favard établit des résultats sur les solutions presque-périodiques au sens de Bohr d'équations différentielles ordinaires linéaires [23]. L'extension de certains résultats

classiques aux équations différentielles stochastiques dans des espaces de Hilbert a suscité un intérêt croissant. Nous considérons deux généralisations des fonctions périodiques : les fonctions S -asymptotiquement périodiques et les fonctions presque automorphes. Le modèle de Leslie est un modèle discret pour tenter de prédire l'évolution d'un nombre d'individus. Si X_n est la taille de la population à l'instant n , le modèle de Leslie s'écrit sous la forme suivante : $X_{n+1} = LX_n$ où L est une matrice appelée la matrice de Leslie et peut être également un opérateur linéaire non borné A dans un espace de Banach \mathbb{X} . Dans ce dernier cas, A peut être considéré comme un générateur infinitésimal d'un semi-groupe. Notre première contribution [73] a été de fournir une condition suffisante d'existence et d'unicité de solution S -asymptotiquement ω -périodique au sens des semi-groupes de l'équation différentielle suivante de type non linéaire à argument constant par morceaux dans \mathbb{X} :

$$\frac{d}{dt}X_t = LX_t + \sum_{j=0}^N A_j(t)X_{[t+j]} + f(t, X_{[t]}), \quad X_0 = \varphi,$$

où $\varphi \in \mathbb{X}$, f est une fonction continue définie sur $\mathbb{R}^+ \times \mathbb{X}$ et $A(t)$ un semi-groupe dépendant du temps exponentiellement stable dans \mathbb{X} . L'étude des équations différentielles à argument constant par morceaux (EPCA) est un sujet important car ces équations ont la structure de systèmes dynamiques continus dans des intervalles de longueur unitaire. Elles combinent donc les propriétés des équations différentielles et des équations aux différences. De nombreux articles ont étudié les EPCA, voir par exemple [13], [14], [15], [16], [17] et les références qui s'y rapportent. L'étude de l'existence de solutions asymptotiquement ω -périodiques est l'un des sujets les plus attrayants de la théorie qualitative en raison de ses applications en biologie mathématique, en théorie du contrôle et en physique. Soit $BC(\mathbb{R}^+, \mathbb{X})$ l'espace de Banach des fonctions continues bornées de \mathbb{R}^+ vers \mathbb{X} ; $P_\omega(\mathbb{R}^+, \mathbb{X})$ le sous espace des fonctions ω -périodiques et $C_0(\mathbb{R}^+, \mathbb{X})$ celui des fonctions nulles à l'infini.

Définition 1 ([96]). *Une fonction $f \in BC(\mathbb{R}^+, \mathbb{X})$ est dite S asymptotiquement ω -périodique si elle peut s'exprimer sous la forme $f = g + h$, avec $g \in P_\omega(\mathbb{R}^+, \mathbb{X})$ and $h \in C_0(\mathbb{R}^+, \mathbb{X})$. L'ensemble de ces fonctions sera désignée par $AP_\omega(\mathbb{R}^+, \mathbb{X})$.*

Le concept de solutions S -asymptotiquement ω -périodiques a été proposé par la suite dans [72] pour une équation dérivée partielle stochastique semi linéaire de la forme suivante :

$$\begin{cases} dX_t = AX_t dt + f(t, X_t)dt + g(t, X_t)dB_t, \\ X_0 = \eta, \end{cases} \quad (3.1)$$

où $\eta \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ et A est un générateur infinitésimal qui génère un C_0 -semigroupe noté $(T(t))_{t \geq 0}$. De plus

$$f : \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \text{ et } g : \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$$

sont continues bornées et $B = (B_t)_{t \geq 0}$ est un mouvement brownien standard bilatéral, défini sur l'espace de probabilité complet filtré $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$. On considère dans la suite des opérateurs A générant un C_0 -semigroupe $T(t)$ exponentiellement stable $\|T(t)\| \leq Me^{-at}$, $a > 0$. Ces équations sont particulièrement importantes par exemple, dans le cadre des équations de la chaleur stochastique avec des conditions limites de type Dirichlet :

$$\begin{cases} dX(t, \eta) = \Delta X(t, \eta)dt + f(X(t, \eta))dt + f(X(t, \eta))dB(t, \eta), & t > 0 \\ X(0, \eta) = h(\eta), & \eta \in D \\ X(t, \eta) = 0, & \eta \in \partial D, \end{cases}$$

dans un domaine borné D et $h(\eta) \in \mathbb{H}$; Δ étant l'opérateur Laplacien.

Théorème 3.1 (Manou-Abi S.M. et Dimbour W. (2019) [72]). *Soit f et g des fonctions limites périodiques en temps $t > 0$ et uniformément bornées sur les sous ensembles de $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$. Supposons de plus qu'il existe des constantes $L_f > 0$ et $L_g > 0$ telles que $\forall t \geq 0, \forall X, Y \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$:*

$$\mathbb{E}\|f(t, X) - f(t, Y)\|^2 \leq L_f \mathbb{E}\|X - Y\|^2 \quad \mathbb{E}\|g(t, X) - g(t, Y)\|^2 \leq L_g \mathbb{E}\|X - Y\|^2.$$

Si

$$2M^2 \left(L_f \frac{1}{a^2} + L_g \frac{1}{a} \right) < 1$$

alors il existe une unique solution S -asymptotiquement ω -périodique au sens des semi-groupes pour l'EDS (3.1).

Ce dernier résultat a nécessité l'introduction de la notion de processus dit limite ω -périodique et l'étude de ses propriétés. Et généralise le concept de fonction limite périodique dans le cas déterministe introduit par Xie et Zhang [21].

Définition 2 (Manou-Abi S.M and Dimbour W. (2019) [72]). *Un processus stochastique continu borné $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ est dite limite ω -périodique s'il existe $\omega > 0$ tel que $\lim_{n \rightarrow +\infty} \mathbb{E}\|X_{t+n\omega} - \tilde{X}_t\|^2 = 0$ pour tout $t \geq 0$ et pour tout $n \in \mathbb{N}$ pour un certain processus stochastique $\tilde{X} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.*

Nous appliquons ensuite les propriétés de tels processus pour établir l'existence et l'unicité de solutions S asymptotiquement ω -périodiques de l'équation (3.1). Une extension de ce résultat afin de généraliser les travaux de [19] au cas d'équations integro-différentielles fractionnaires stochastiques dirigées par un mouvement brownien a été établi dans [82] :

$$dX_t = \int_0^t \frac{(t-s)^{\eta-2}}{\Gamma(\eta-1)} AX_s ds dt + g(t, X_t) dB_t, \quad X_0 = c_0 \quad (3.2)$$

avec $\eta \in (1, 2)$ et $\Gamma(\cdot)$ désignant la fonction Gamma. L'intégrale convoluée dans (3.2) est connue sous le nom d'intégrale fractionnaire de Riemann-Liouville. Dans le cadre des équations intégro-différentielles stochastiques dirigées par des processus de Lévy à valeurs dans un espace de Hilbert \mathbb{H} , nous avons étudié dans [68], l'existence de solutions presque automorphe en loi et en moyenne quadratique sous certaines hypothèses appropriées de la classe suivante :

$$\begin{aligned} dX_t &= (AX_t + g(t, X_t))dt + \int_{-\infty}^t B_1(t-s)f(s, X_s)ds \\ &+ \int_{-\infty}^t B_2(t-s)h(s, X_s)dW_s \\ &+ \int_{-\infty}^t B_2(t-s) \int_{|y|_V < 1} F(s, X_{s-}, y)\tilde{N}(ds, dy) \\ &+ \int_{-\infty}^t B_2(t-s) \int_{|y|_V \geq 1} G(s, X_{s-}, y)N(ds, dy) \end{aligned} \quad (3.3)$$

où $A : D(A) \subset H$ est le générateur infinitésimal d'un C_0 -semigroupe $(T(t))_{t \geq 0}$, B_1 et B_2 sont des noyaux de convolution dans $\mathbb{L}^1(0, \infty)$ et $\mathbb{L}^2(0, \infty)$ respectivement. $f, g : \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$; $h : \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow L(V, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, $F, G : \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \times V \rightarrow$

$L^2(\mathbb{P}, \mathbb{H})$; W et N sont les composantes (pour le Brownien et Poisson) de la décomposition de Lévy-Itô du processus de Lévy bilatéral L avec $(H, \|\cdot\|)$ and $(V, |\cdot|)$ des espaces de Hilbert séparables. Soit $SBC(\mathbb{R}, L^2(\mathbb{P}, \mathbb{H}))$ l'espace des processus stochastiquement continus et bornés (espace de Banach) muni de la norme $\|X\| = \sup_{t \in \mathbb{R}} (\mathbb{E}\|X_t\|^2)^{\frac{1}{2}}$. Sur la base du théorème du point fixe de Schauder, l'existence d'une solution presque automorphe en loi et en moyenne quadratique a été établie en utilisant des conditions plus faibles que les conditions de Lipschitz.

3.2 Approximation des solutions et processus stables

Dans le cadre des méthodes d'approximation numérique, la simulation des solutions est bien connue dans la littérature pour de nombreux processus de Lévy. L'analyse numérique de ces EDS se concentre essentiellement sur les schémas d'approximation en temps discret, l'approximation des trajectoires ou l'approximation des espérances de la solution, etc. Elle est développée depuis de nombreuses années. Le schéma d'Euler-Maruyama est une méthode bien connue pour l'approximation numérique des solutions lorsqu'elles existent. Dans [86], nous avons étudié le problème des approximations numériques afin de contribuer à enrichir la littérature existante et récente à ce sujet dans le cadre des équations différentielles stochastiques dirigées par des processus de Lévy α -stables :

$$\begin{cases} dX_t = f(X_t)dt + \phi(X_{t-})dZ_t, & t \in [0, T] \\ X_0 = x_0 \in \mathbb{R}, \end{cases} \quad (3.4)$$

où x_0 est une valeur réelle initiale, $T > 0$ un horizon de temps, et les fonctions f, ϕ vérifiant certaines conditions de croissance linéaire. Le processus Z est un processus stable défini comme suit.

Définition 3 ([103]). *Un processus stochastique $(Z_t)_{t \in [0, T]}$ défini sur un espace de probabilité filtré et complet $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0, T]}, \mathbb{P})$ est un processus α -stable de caractéristiques (b, c_+, c_-) , s'il est un processus de Lévy de fonction caractéristique donnée par :*

$$\varphi_{Z_t}(u) = \exp t \left(iub + \int_{-\infty}^{+\infty} (e^{iuy} - 1 - iuy\mathbb{I}_{|y| \leq 1})\nu(dy) \right), \quad t \in [0, T],$$

où b représente le paramètre de dérive de Z et ν la mesure de Lévy définie sur $\mathbb{R} \setminus \{0\}$ par

$$\nu(dx) := \frac{dx}{|x|^{\alpha+1}} \left(c_+ 1_{\{x>0\}} + c_- 1_{\{x<0\}} \right).$$

Les paramètres c_+, c_- ci-dessus sont non négatifs avec $c_+ + c_- > 0$ et $c_+ = c_-$ lorsque $\alpha = 1$. Le processus est dit symétrique si $c_+ = c_- := c$. Il est dit strictement α -stable si $b = 0$. On peut également définir ces processus à partir d'un processus de Lévy et des lois α -stables. Il faut noter qu'il existe plusieurs caractérisations des lois α -stables et que chacune d'entre elles présente des avantages et inconvénients en fonction de l'objectif poursuivi.

Définition 4 ([108]). *Une variable aléatoire X suit une distribution α -stable avec $\alpha \in (0, 2]$ et on note $X \sim S_\alpha(\gamma, \beta, \zeta; 1)$ si elle est déterminée de manière unique par sa fonction*

caractéristique :

$$\Psi(t) = \mathbb{E}(\exp(itX)) = \begin{cases} \exp\left(-\gamma^\alpha |t|^\alpha \left[1 - i\beta \left(\tan\left(\frac{\pi\alpha}{2}\right)\right) \operatorname{sgn}(t)\right] + i\zeta t\right) & \text{if } \alpha \neq 1. \\ \exp\left(-\gamma |t| \left[1 + i\beta \frac{2}{\pi} \operatorname{sgn}(t) \log(|t|)\right] + i\zeta t\right) & \text{if } \alpha = 1 \end{cases}$$

avec $\alpha \in (0, 2]$, $\beta \in [-1, 1]$, $\gamma > 0$, $\zeta \in \mathbb{R}$ et $\operatorname{sgn}(t)$ étant la fonction signe.

La définition ci-dessus correspond à la paramétrisation 1. La définition suivante est la paramétrisation 0 et est utile pour les aspects numériques. La caractérisation des lois stables fait ressortir quatre types de paramètres.

- L'indice de stabilité $\alpha \in (0, 2]$: il s'agit d'un coefficient d'aplatissement qui caractérise la queue de distribution.
- Le paramètre d'échelle $\gamma > 0$, un paramètre de dispersion : plus il est important, plus les courbes sont volatiles. Pour $\gamma = 1$ on dit que X est de loi standard α -stable.
- Le paramètre $\beta \in [-1, 1]$ est le paramètre d'asymétrie. Lorsqu'il est nul, la distribution est symétrique par rapport au paramètre de position ζ . Pour $\alpha \neq 1$ et $\zeta = 0$ on dit que X est de loi strictement α -stable. Si de plus $\beta = 0$ alors on dit que X est symétrique.

Définition 5 ([34]). *Une variable aléatoire X suit une distribution α -stable avec $\alpha \in (0, 2]$ et on note $X \sim S_\alpha(\gamma, \beta, \zeta; 0)$ si elle est déterminée de manière unique par sa fonction caractéristique :*

$$\phi(t) = \begin{cases} \exp\left(-\gamma^\alpha |t|^\alpha \left[1 + i\beta \left(\tan\left(\frac{\pi\alpha}{2}\right)\right) \operatorname{sgn}(t) (|\gamma t|^{1-\alpha} - 1)\right] + i\zeta t\right) & \text{if } \alpha \neq 1. \\ \exp\left(-\gamma |t| \left[1 + i\beta \frac{2}{\pi} \operatorname{sgn}(t) \log(\gamma |t|)\right] + i\zeta t\right) & \text{if } \alpha = 1. \end{cases}$$

Ces différents paramétrisations ont souvent donné lieu à des confusions dans la littérature. Nolan discute également de la paramétrisation à choisir dans [34]. La première paramétrisation proposée par [108] n'assure pas la continuité de la fonction de densité lorsque $\alpha = 1$ et $\beta = 0$ (à cause du terme $\tan(\pi\alpha/2)$), Elle ne fournit pas non plus une famille de localisation d'échelle lorsque $\alpha = 1$ (à cause du terme $\gamma \log(\gamma)$), alors que la seconde paramétrisation est continue par rapport à tous les paramètres. Les deux formulations ci-dessus sont liées par l'équation clé suivante

$$\zeta = \begin{cases} \mu + \beta\gamma \tan\left(\frac{\pi\alpha}{2}\right) & \text{if } \alpha \neq 1 \\ \mu + \beta \frac{2}{\pi} \gamma \log(\gamma) & \text{if } \alpha = 1. \end{cases}$$

Il convient de noter que les distributions α -stables ont une variance infinie pour tout $\alpha < 2$, et de moyenne infinie (indéfinie) pour $\alpha \in (0, 1]$, ce qui rend l'estimation des paramètres difficile. Notons que la moyenne, si elle existe ($\alpha > 1$), est la mesure naturelle de la localisation, mais elle ne peut pas être estimée aussi précisément que ζ .

Définition 6 ([103]). *Un processus stochastique \mathcal{F}_t -adapté $Z = \{Z_t\}_{t \geq 0}$ est un processus α -stable process si*

1. $Z_0 = 0$, p.s. ;
2. Z a des incréments stationnaires de loi α -stable : $Z_t - Z_s \sim Z_{t-s} \sim S_\alpha((t-s)^{1/\alpha} \sigma, \beta, \omega; 1)$ ou $S_\alpha((t-s)^{1/\alpha} \sigma, \beta, \omega; 0)$, $t > s \geq 0$;
3. Pour tout instant $0 \leq s_0 < \dots < s_m < \infty$, les variables aléatoires $Z_{s_0}, Z_{s_1} - Z_{s_0}, \dots, Z_{s_m} - Z_{s_{m-1}}$ sont indépendants.

L'algorithme suivant issu de [51] permet de simuler des processus strictement α -stables. Des transformations algébriques permettent de l'étendre au cas général.

Algorithm 1 Simulation des trajectoires d'un processus strictement stable

- 1: Soit $Z = (Z_t)_{t \in [0, T]}$ un processus strictement α -stable.
- 2: **Etape 1 :**
- 3: Simuler n variables aléatoires indépendantes et uniformément distribuées Φ sur $[-\pi/2, \pi/2]$ et n variables aléatoires indépendantes et identiquement distribuées W de loi exponentielle de paramètre 1.
- 4: **Etape 2 :** Calculer ΔZ_i pour tout $i = 1, \dots, n$ comme suit.
 1. Si $\alpha \neq 1$

$$\Delta Z_i = \sigma \left(\frac{T}{n} \right)^{1/\alpha} \frac{\sin(\alpha(\Phi - \phi_0))}{\cos(\Phi)^{1/\alpha}} \left(\frac{\cos(\Phi - \alpha(\Phi - \phi_0))}{W} \right)^{\frac{1-\alpha}{\alpha}}.$$

2. Si $\alpha = 1$

$$\Delta Z_i = \sigma \left(\frac{T}{n} \right)^{1/\alpha} \frac{2}{\pi} \left(\left(\frac{\pi}{2} + \beta\Phi \right) \tan(\Phi) - \beta \log \left(\frac{\frac{1}{2}\pi W \cos(\Phi)}{\frac{1}{2}\pi + \beta\Phi} \right) \right)$$

avec

$$\phi_0 = -\frac{\beta\pi}{2} \frac{1 - |1 - \alpha|}{\alpha}.$$

- 5: **Etape 3 :** La trajectoire discrétisée de Z est donnée par

$$Z(t_i) = \sum_{k=1}^i \Delta Z_k.$$

D'un point de vue pratique, nous considérons la version linéaire interpolée entre les instants de discrétisation. On rappelle que le processus Ornstein-Uhlenbeck dirigé par un processus α -stable $(Z_t)_{t \geq 0}$ est défini par

$$dX_t = -\theta X_t dt + \rho dZ_t, \quad X_0 = x_0 \in \mathbb{R}. \quad (3.5)$$

et admet la représentation intégrale

$$X_t(\omega) = e^{-\theta t} \left(x_0 + \rho \int_0^t e^{\theta s} dZ_s(\omega) \right), \quad \text{for } t \in [0, T].$$

Un processus de type Cox–Ingersoll–Ross process dirigé par un processus α -stable $(Z_t)_{t \geq 0}$ peut être défini comme suit :

$$dX_t = (\lambda - \theta X_t) dt + \rho |X_t|^q dZ_t, \quad X_0 \geq 0, \quad (3.6)$$

avec $q \in [0, 1)$, λ , ρ et θ des constantes. On peut aussi considérer les modèles de type Lotka et de Volterra utilisés en dynamique des populations. Un processus de type Lotka–Volterra dirigé par un processus α -stable $(Z_t)_{t \geq 0}$ peut être défini comme suit :

$$dX_t = X_t(\lambda - \theta X_t) dt + \rho |X_t|^q dZ_t, \quad X_0 \geq 0,$$

avec $q \in (0, 1]$, λ , ρ et θ des constantes. Une solution approximative d'Euler-Maruyama de l'équation (3.4) est donnée comme suit :

$$\tilde{X}_t^n = x_0 + \int_0^t f(\tilde{X}_{\pi_s^n}^n) ds + \int_0^t \phi(\tilde{X}_{\pi_s^n}^n) dZ_s, \quad (3.7)$$

avec

$$\pi_t^n = \frac{k}{n} \quad \text{si } t \in \left[\frac{k}{n}, \frac{(k+1)}{n} \right], \quad k = 0, \dots, n-1.$$

Quelques exemples de trajectoires simulées de processus stables et d'EDS sont présentées aux Figures (5.1) et (5.2). J'ai étudié dans [86] cette classe ci-dessus d'équations différentielles stochastiques dirigées par des processus α -stables sous les conditions particulières suivantes pour tout $p > 0$, $x \in \mathbb{R}$:

$$\text{(C1)} \quad |\phi(x)|^p \leq L_\phi(1 + |x|^p) \quad \text{(C2)} \quad |f(x)|^p \leq L(1 + |x|^p) L_\phi, L > 0.$$

Rappelons un résultat d'existence faible de [105].

Lemme 3.1 ([105]). *Soit Z un processus α -stable avec $\alpha \in (0, 2)$ and $\alpha \neq 1$. Soit f et ϕ des fonctions continues vérifiant (C1) et (C2). Il existe une solution faible pour l'EDS (3.4) telle que pour tout $p \in (0, \alpha)$, on ait*

$$\sup_{t \in [0, T]} \mathbb{E}|X_t|^p < \infty.$$

Des exemples particuliers des conditions (C1) et (C2) sont les fonctions continues Lipschitz et Hölder.

L'existence de solutions approximative du schéma d'Euler-Maruyama est prouvée sous certaines conditions. Par exemple, lorsque f est continue bornée et Hölérienne, ϕ continue et Lipschitzienne, il a été établi dans [104] des résultats d'approximation lorsque Z est un processus α -stable symétrique dans \mathbb{R}^d .

Nous proposons dans [86] un nouveau résultat d'approximation dans le cas où Z est un processus α -stable réel et pas nécessairement symétrique basé sur une approche différente. La méthodologie est nouvelle et le résultat également nouveau. La technique consiste à déployer convenablement une méthode de troncature en séparant les sauts du processus stable via la décomposition de Lévy-Itô de Z :

$$Z_t = b_R t + \int_0^t \int_{|x| \leq R} x (\mu - \sigma)(ds, dx) + \int_0^t \int_{|x| > R} x \mu(ds, dx), \quad t \in [0, T],$$

où R est un niveau de troncature positif arbitraire (classiquement choisi à 1) et μ est la mesure aléatoire de Poisson gouvernant les sauts du processus sur $[0, T] \times \mathbb{R}$ avec intensité $\sigma(dt, dx) = dt \otimes \nu(dx)$; et b_R étant le paramètre de dérive donné par

$$b_R := b + \int_{1 < |x| \leq R} x \nu(dx).$$

Pour établir la convergence de \tilde{X}^n vers X , nous partons de l'expression explicite suivante

$$\tilde{X}_t^n - X_t = \int_0^t [f(\tilde{X}_{\pi_s^n}^n) - f(X_s)] ds + \int_0^t [\phi(\tilde{X}_{\pi_s^n}^n) - \phi(X_{\pi_s^n})] dZ_s + \int_0^t [\phi(X_{\pi_s^n}) - \phi(X_s)] dZ_s,$$

en considérant la décomposition de Lévy-Itô de Z , un choix de troncature adéquat, des techniques de calcul stochastique et le Lemme de Gronwall.

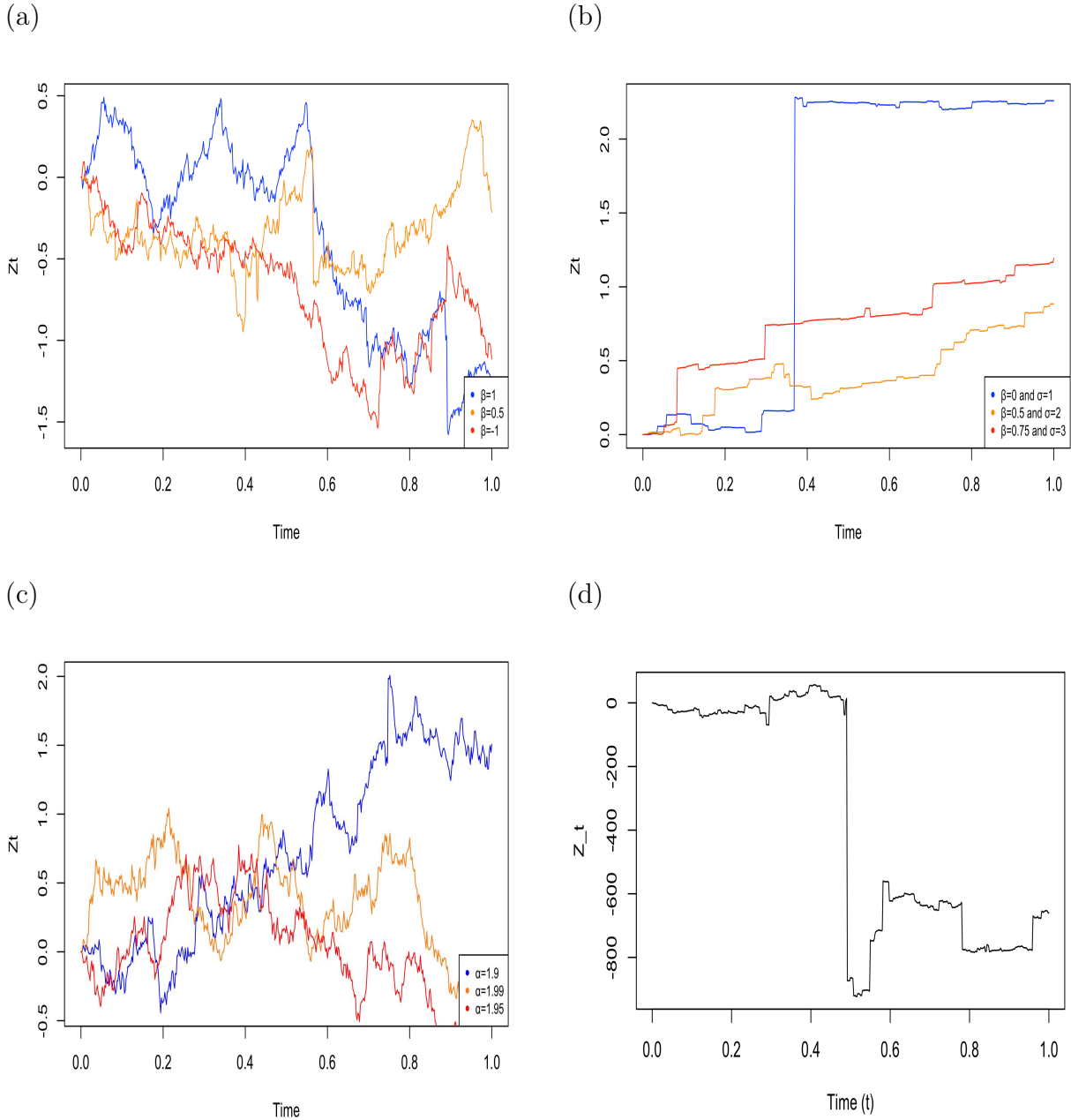


FIGURE 3.1 – Trajectoires d’un processus standard strictement 1.7-stable avec différentes asymétries (a), d’un processus 0.7-stable avec différentes échelles et asymétries (b), du processus symétrique et standard avec différents indices de stabilité α (c) et le cas d’un processus standard de Cauchy 1-stable (d).

Théorème 3.2 (Manou-Abi S.M. [86] **En révision**). *Soit Z un processus α -stable avec $\alpha \in (1, 2)$ et $(X_t)_{t \in [0, T]}$ une solution de l’EDS (3.4) tel que $X_0 = x_0$ presque sûrement. Sous les conditions (C1), (C2) et $1 \leq 2p < \alpha$, il existe une constante positive C dépendant des paramètres $L, p, \alpha, L_\phi, b, c_+, c_-$ telle que*

$$\mathbb{E} \sup_{t \in [0, T]} |\tilde{X}_t^n - X_t|^{2p} \leq C n^{-p^2/2}.$$

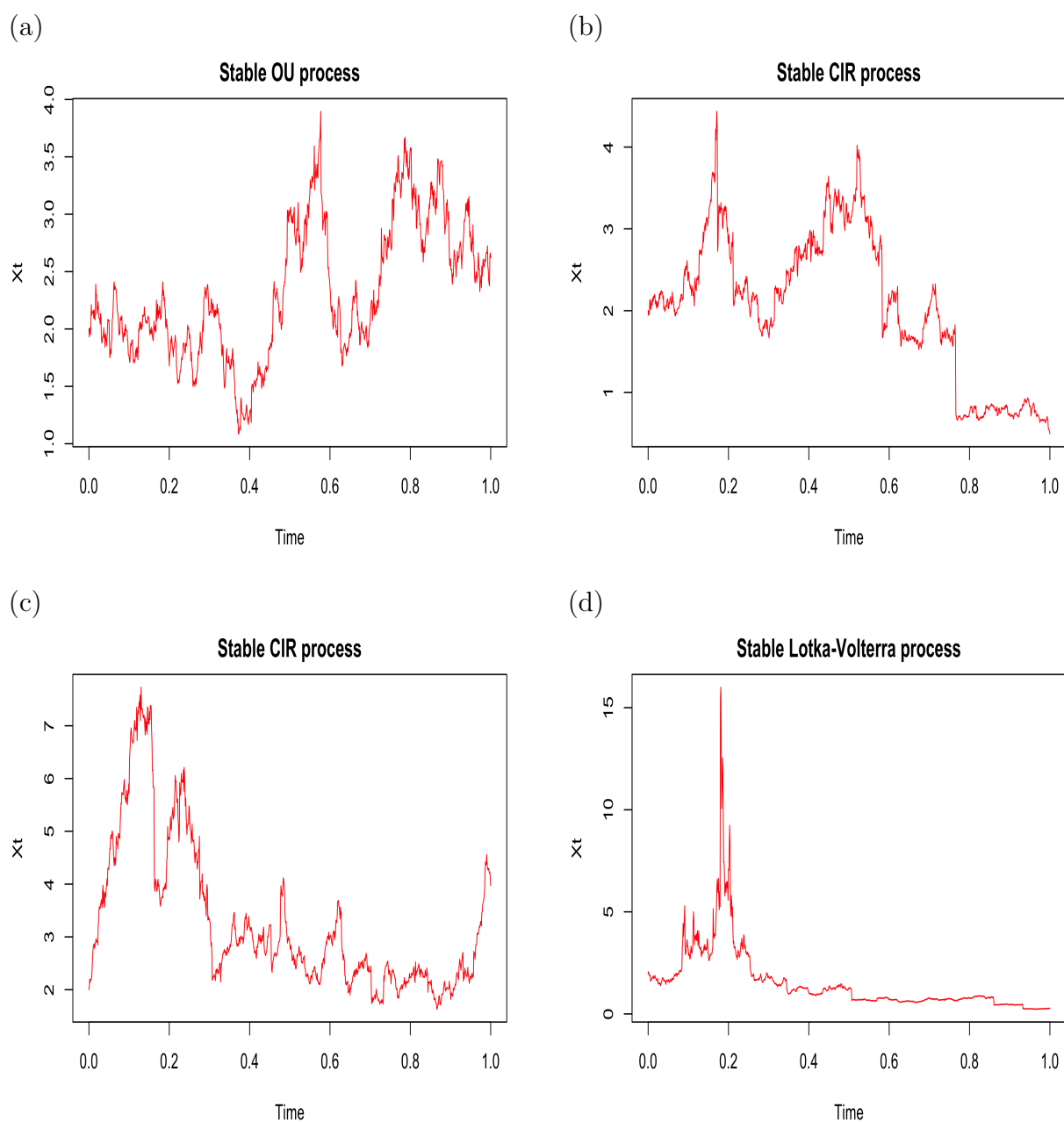


FIGURE 3.2 – Trajectoires approximatives d'un processus OU standard strictement stable avec $n = 1000$, $\rho = 1$, $\theta = 1.2$, $\alpha = 1.7$ et $\beta = 0$ (a); d'un processus CIR standard strictement stable avec $n = 1000$, $\rho = 1$, $\theta = 1.2$, $\alpha = 1.7$, $\beta = 0$, $\lambda = 0.2$ et $q = \alpha^{-1}$ (b); d'un processus CIR standard strictement stable avec $n = 1000$, $\rho = 1$, $\theta = 1.2$, $\alpha = 1.7$, $\beta = 0$, $\lambda = 0.2$ et $q = 1 - \alpha^{-1}$ (c) et le cas d'un processus de Lotka-Volterra standard strictement stable avec $\lambda = 0.2$, $\theta = 1.2$, $\alpha = 1.7$, $\beta = 0$, $\rho = 1$, $q = 1/2$ et $n = 1000$ (d).

Statistique des processus stochastiques

Je considère ici le problème d'estimation des paramètres de processus stochastiques issus de modèles markoviens et des solutions d'équations différentielles stochastiques dirigées par des processus α -stables. Les publications relatives à ces travaux se retrouvent dans [87] et [93].

4.1 Un modèle markovien pour étudier la dynamique de transmission de la fièvre typhoïde

Un modèle probabiliste a été étudié dans [93] pour essayer de capturer la dynamique de transmission de la fièvre typhoïde à Mayotte en utilisant une approche markovienne. La fièvre typhoïde fait partie des maladies hydriques (transmission passant par la contamination de l'eau) causées par un contact direct ou indirect avec un individu infecté ou avec l'environnement (nourriture, contamination de l'eau). On peut décrire la dynamique de manière simplifiée comme suit. Chaque individu en contamine un nouveau avec un taux d'infection (de naissance) constant λ , indépendamment des autres individus. Les individus infectés se rétablissent indépendamment les uns des autres à un taux de guérison (décès) constant μ . Les nouveaux cas exogènes arrivent à un taux (d'immigration) constant ν , également indépendamment des autres contaminations et guérisons. Nous décrivons le modèle stochastique de dénombrement des personnes infectées comme suit. Soit X_t le nombre d'individus infectés au temps $t \geq 0$. Le processus de comptage $X = (X_t)_{t \geq 0}$ est alors un processus linéaire de naissance et de mort avec immigration (processus LBDI). Il appartient à la classe des processus de saut de Markov. La distribution de X est caractérisée par son générateur infinitésimal, la matrice Q à entrées non nulles donnée par

$$\begin{aligned} Q(i, i + 1) &= \lambda i + \nu && \text{pour } i \geq 0, \\ Q(i, i - 1) &= \mu i && \text{pour } i \geq 1, \\ Q(i, i) &= -(\lambda + \mu)i - \nu && \text{pour } i \geq 0. \end{aligned}$$

Le semi-groupe de transition est défini par $(P(t) = (p_{i,j}(t), i, j \in \mathbb{N}))_{t \geq 0}$, où $p_{i,j}(t) = \mathbb{P}(X_t = j | X_0 = i)$, $i, j \in \mathbb{N}, t \geq 0$. Les trajectoires de $X = (X_t)_{t \geq 0}$ sont à pas constants et ponctuées de sauts d'amplitude $+1$ ou -1 . Le processus X est connu sous l'appellation processus de naissance et de mort linéaire avec immigration (LBDI en anglais). La simulation de ce processus permet de capturer trois régimes possibles : croissance exponentielle

(régime transcient) si $\lambda > \mu$, tous les états sont visités une infinité de fois avec un temps de retour moyen infini (régime récurrent nul) si $\lambda = \mu$, et tous les états sont visités une infinité de fois avec un temps de retour moyen fini (régime récurrent positif) si $\lambda < \mu$. Dans ce dernier cas il existe une unique mesure invariante $\pi = (\pi_i, i \in \mathbb{N})$ donnée par

$$\pi_i = \binom{r+i-1}{i} \left(\frac{\lambda}{\mu}\right)^i \left(1 - \frac{\lambda}{\mu}\right)^r, \quad (4.1)$$

avec $r = \nu/\lambda$. L'objectif visé est l'estimation des paramètres (λ, μ, ν) dans un régime récurrent positif. De plus, la principale originalité et difficulté de cette étude vient du cadre d'observation. En effet, les effectifs de la population infectée sont cachés. Les seules données disponibles sont les cas cumulés (en jour ou semaine) des nouveaux infectés. Ceci rend difficile l'estimation des paramètres. Nous dérivons d'abord une expression analytique des paramètres inconnus en tant que fonctions de probabilités des transitions à temps discrets suffisamment observés. Par exemple, en considérant les transitions du type $p_{0,0}, p_{0,1}, p_{1,0}$, nous obtenons :

Théorème 4.1 (Bouzalmat I., De Saporta B. et Manou-Abi S.M. [93]). *Les paramètres λ, μ et ν sont donnés par la relation suivante : $(\lambda, \mu, \nu) = g(p_{0,0}, p_{0,1}, p_{1,0})$ avec*

$$\begin{aligned} \lambda &= g_1(p_{0,0}, p_{0,1}, p_{1,0}) = \frac{\ln\left(\frac{u}{q}\right)(q-1)}{\Delta t(q-u)} \\ \mu &= g_2(p_{0,0}, p_{0,1}, p_{1,0}) = \frac{\ln\left(\frac{u}{q}\right)(u-1)}{\Delta t(q-u)} \\ \nu &= g_3(p_{0,0}, p_{0,1}, p_{1,0}) = \frac{\ln(p_{0,0}) \ln\left(\frac{u}{q}\right)(q-1)}{\ln(q) \Delta t(q-u)}, \end{aligned}$$

où

$$q = \frac{p_{0,1}}{p_{0,0} \ln(p_{0,0})} W \left(\frac{\left(\frac{p_{0,0}}{p_{0,1}} + 1\right) \ln(p_{0,0})}{p_{0,1}} \right),$$

$$u = 1 - \frac{p_{1,0}}{p_{0,0}} : \text{ et } W \text{ désignant ici la fonction de Lambert [57].}$$

On obtient ainsi des estimateurs consistents et asymptotiquement de lois gaussiennes dès lors qu'on dispose d'estimateurs consistents des probabilités de transition dans un régime récurrent positif.

Théorème 4.2 (Bouzalmat I., De Saporta B. et Manou-Abi S.M. [93]). *Soit $\hat{p}_{0,0}^n, \hat{p}_{0,1}^n$ et $\hat{p}_{1,0}^n$ des estimateurs consistents de $p_{0,0}, p_{0,1}$ et $p_{1,0}$, asymptotiquement de loi normale de matrice de variance-covariance Σ' au sens suivant :*

$$\sqrt{n} \begin{pmatrix} \hat{p}_{0,0}^n - p_{0,0} \\ \hat{p}_{0,1}^n - p_{0,1} \\ \hat{p}_{1,0}^n - p_{1,0} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}_3(0, \Sigma').$$

avec

$$\Sigma' = \begin{pmatrix} \frac{p_{0,0}(1-p_{0,0})}{\pi_0} & -\frac{p_{0,0}p_{0,1}}{\pi_0} & 0 \\ -\frac{p_{0,0}p_{0,1}}{\pi_0} & \frac{p_{0,1}(1-p_{0,1})}{\pi_0} & 0 \\ 0 & 0 & \frac{p_{1,0}(1-p_{1,0})}{\pi_1} \end{pmatrix}, \quad (4.2)$$

et π étant la loi invariante mentionnée dans (4.1). Alors les estimateurs $(\hat{\lambda}^n, \hat{\mu}^n, \hat{\nu}^n) = g(\hat{p}_{0,0}^n, \hat{p}_{0,1}^n, \hat{p}_{1,0}^n)$ issus du Théorème 4.1 sont aussi consistents et asymptotiquement de loi normal de matrice de variance-covariance $\Sigma = Dg\Sigma'Dg^T$, où Dg est la matrice Jacobienne issue de l'application g et Dg^T sa transposée.

Considérons le cas où l'on observe complètement les transitions et soit $\hat{p}_{i,j}^n$ l'estimateur du maximum de vraisemblance de $p_{i,j}$ sur des observations $(X_{k\Delta t}, 0 \leq k \leq n)$ et des transitions du type $(i, j) \in \{(0, 0), (0, 1), (1, 0)\}$. Ces estimateurs correspondent au rapport du nombre de transitions observées de i à j sur le nombre total de transitions observées partant de i dans la séquence $(X_{k\Delta t}, 0 \leq k \leq n)$:

$$\hat{p}_{i,j}^n = \frac{\sum_{k=0}^{n-1} 1_{\{X_k=i, X_{k+1}=j\}}}{\sum_{k=0}^{n-1} 1_{\{X_k=i\}}}.$$

Nous illustrons la performance de ces estimateurs en simulant le modèle ci-dessus avec $X_0 = 0$, $\lambda = 0.03$, $\mu = 0.1$ et $\nu = 0.01$. Les valeurs estimées de ces paramètres sont données comme suit dans le tableau 4.1, y compris l'erreur asymptotique (divisée par \sqrt{n}) dans le tableau 4.2.

Le tableau 4.1 montre que la propriété de convergence est satisfaite, bien qu'un nombre élevé d'observations soit nécessaire pour capturer le bon ordre de grandeur. Cela confirme que même dans un cadre d'observation complet, la récupération des paramètres à partir d'observations discrètes du processus est exigeante. Le tableau 4.2 montre que la variance asymptotique des estimateurs dépend fortement du pas de temps et augmente à mesure que ce dernier diminue ; l'impact étant plus fort sur le paramètre taux d'immigration μ . Cependant, dans le cadre des données réelles que nous manipulons, la séquence $(X_{k\Delta t}, 0 \leq k \leq n)$ contient des valeurs cachées. Nous considérons donc que X_t n'est pas complètement observé. Seuls les nouveaux cas cumulés d'isolés sur des laps de temps donnés sont disponibles. Bien que cette formulation soit très courante dans les données de santé publique, elle est mathématiquement originale et difficile. Par conséquent, il conviendrait d'utiliser d'autres approches d'estimation des probabilités de transition $p_{0,0}$, $p_{0,1}$ et $p_{1,0}$ à partir des outils de la théorie des chaînes de Markov cachées. Soit Δt un pas de temps fixé (peut être un jour ou une semaine). Soit $Y_{(a,b]}$ le cumul des nouveaux cas isolés correspondant au processus X sur l'intervalle de temps $(a, b]$.

Le processus discret joint $(X_{n\Delta t}, Y_{(n\Delta t, (n+1)\Delta t]})_{n \in \mathbb{N}}$ est une chaîne de Markov cachée (HMM en anglais). La propriété de Markov nous permet d'obtenir les probabilités d'émission de $(Y_{(n\Delta t, (n+1)\Delta t]})_{n \in \mathbb{N}}$ conditionnellement à $(X_{n\Delta t}, 0 \leq k \leq n)$. Soit $Z_n = (X_{n-1}, X_n)$ la chaîne de Markov cachée bi-dimensionnelle à valeurs dans l'espace d'état \mathbb{N}^2 . On montre tout d'abord que (Z_n, Y_n) est une chaîne de Markov cachée possédant les caractéristiques suivantes. Le processus $(Z_n, Y_n)_{n \in \mathbb{N}^*}$ est une chaîne de Markov cachée dont les caractéristiques sont données par le triplet $M = (Q, \psi, \rho)$, où

- la matrice de transition Q de la chaîne de Markov cachée (Z_n) est

$$Q_{(i,j),(i',j')} = \mathbb{P}(Z_{n+1} = (i', j') | Z_n = (i, j)) = p_{i',j'} \delta_{i=j},$$

TABLE 4.1 – Valeurs estimées et erreur standard lorsque les nombres d’infectés sont observés avec une période Δt et un horizon de temps H .

		$H = 1000$		
Pas de temps Δt		1	7	30
Nombre d’observations n		1001	143	34
$\hat{\lambda}^n$		0.027 (0.100)	0.081 (0.053)	0.049 (0.065)
$\hat{\mu}^n$		0.084 (0.035)	0.105 (0.046)	0.070 (0.128)
$\hat{\nu}^n$		0.012 (0.484)	0.007 (0.106)	0.007 (0.076)
		$H = 5000$		
Pas de temps Δt		1	7	30
Nombre d’observations n		5001	715	167
$\hat{\lambda}^n$		0.041 (0.045)	0.052 (0.024)	0.064 (0.029)
$\hat{\mu}^n$		0.088 (0.016)	0.132 (0.020)	0.117 (0.058)
$\hat{\nu}^n$		0.009 (0.216)	0.012 (0.047)	0.009 (0.034)
		$H = 50000$		
Pas de temps Δt		1	7	30
Nombre d’observations n		50001	7143	1667
$\hat{\lambda}^n$		0.030 (0.014)	0.023 (0.007)	0.020 (0.009)
$\hat{\mu}^n$		0.098 (0.004)	0.097 (0.006)	0.102 (0.018)
$\hat{\nu}^n$		0.010 (0.068)	0.010 (0.014)	0.010 (0.011)

TABLE 4.2 – Variance asymptotique des estimateurs

Pas de temps Δt	1	7	30
$\Sigma_{11}^{1/2} (\lambda)$	3.190	0.632	0.377
$\Sigma_{22}^{1/2} (\mu)$	1.112	0.552	0.749
$\Sigma_{33}^{1/2} (\nu)$	15.309	1.264	0.443

pour tout $(i, j), (i', j') \in \mathbb{N}^2$;

- la probabilité d'émission ψ du processus Y sachant le processus Z est

$$\psi_{(i,j)}(y) = \mathbb{P}(Y_n = y | Z_n = (i, j)) = \frac{p_{i,(j,y)}}{p_{i,j}},$$

pour $i, j, y \in \mathbb{N}$, avec $p_{i,(j,y)} = p_{i,(j,y)}(\Delta t)$ donnée par l'équation précédente (??) ;

- la loi initiale ρ du processus des états Z est

$$\rho_{i,j} = \mathbb{P}(Z_1 = (i, j)) = p_{i,j}\pi_i,$$

pour $i, j \in \mathbb{N}$, avec π la loi stationnaire de X donnée à l'équation (4.1). L'algorithme standard pour estimer les paramètres les plus probables $M = (Q, \psi, \rho)$ compte tenu des observations est l'algorithme forward-backward ou de Baum-Welch, qui est un cas particulier de l'algorithme EM (Expectation-Maximization). Le principal atout de l'algorithme de Baum-Welch est que la vraisemblance peut être maximisée explicitement. Malheureusement, l'exécution de la procédure standard du package HMM sur (Z_n, Y_n) produit une instabilité numérique et des erreurs dues aux nombreuses entrées nulles dans la matrice de transition Q de Z . Nous avons donc ré-écrit la procédure, afin d'obtenir des estimateurs explicites. Ceci nous permet donc d'adapter l'algorithme de Baum-Welch afin d'estimer ces probabilités de transition en temps discret.

Théorème 4.3 (Bouzalmat I., De Saporta B. et Manou-Abi S.M. [93]). *Étant donné les paramètres $M^n = (Q^n, \psi^n, \rho^n)$, les estimations du maximum de vraisemblance de $M^{n+1} = (Q^{n+1}, \psi^{n+1}, \rho^{n+1})$ conditionné aux observations (y_1, \dots, y_T) sont données, pour i, j, y dans \mathbb{N} , par*

$$\begin{aligned} Q_{(i,j),(i',j')}^{n+1} &= \frac{\sum_{t=1}^T \xi_{(i,j),(i',j')}^n(t)}{\sum_{t=1}^T \gamma_{(i,j)}^n(t)} \delta_{i'=j}, \\ \psi_{(i,j)}^{n+1}(y) &= \frac{\sum_{t=1}^T 1_{y_t=y} \gamma_{(i,j)}^n(t)}{\sum_{t=1}^T \gamma_{(i,j)}^n(t)}, \\ \rho_{i,j}^{n+1} &= \gamma_{(i,j)}^n(1), \\ p_{i,j}^{n+1} &= \frac{\sum_{i' \in \mathbb{N}} Q_{(i',i),(i,j)}^{n+1}}{\sum_{i' \in \mathbb{N}} 1_{Q_{(i',i),(i,j)}^{n+1} \neq 0}}, \end{aligned}$$

avec

$$\begin{aligned} \gamma_{(i,j)}^n(t) &= \frac{\alpha_{(i,j)}^n(t) \beta_{(i,j)}^n(t)}{\sum_{i,j \in \mathbb{N}} \alpha_{(i,j)}^n(t) \beta_{(i,j)}^n(t)}, \\ \xi_{(i,j),(i',j')}^n(t+1) &= \frac{\alpha_{(i,j)}^n(t) p_{(i',j')}^n \psi_{(i',j')}^n(y_{t+1}) \beta_{(i',j')}^n(t+1)}{\sum_{i,j \in \mathbb{N}} \alpha_{(i,j)}^n(t) \beta_{(i,j)}^n(t)} \delta_{i'=j}, \\ p_{i,j}^n &= \frac{\sum_{i' \in \mathbb{N}} Q_{(i',i),(i,j)}^n}{\sum_{i' \in \mathbb{N}} 1_{Q_{(i',i),(i,j)}^n \neq 0}}, \end{aligned}$$

et les suites progressives et rétrogrades α^n et β^n sont définies pour i, j dans \mathbb{N} et $2 \leq t \leq T$

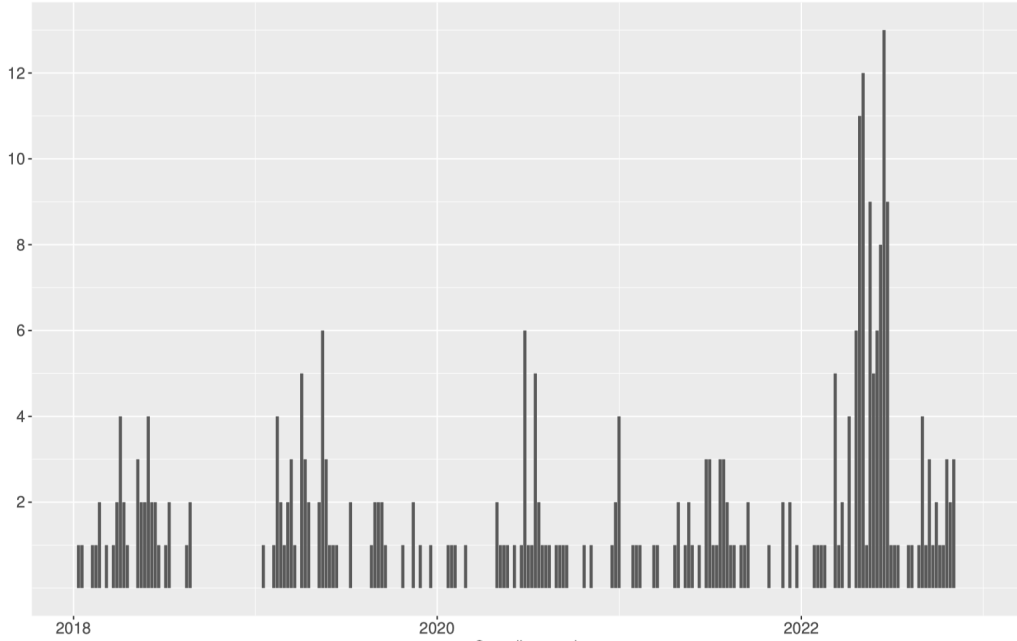


FIGURE 4.1 – Nombre de nouveaux cas confirmés déclarés cumulés hebdomadaires de fièvre typhoïde à Mayotte entre 2018 et 2022.

TABLE 4.3 – Proportions des transitions quotidiennes observées dans les cas déclarés de fièvre typhoïde à Mayotte entre 2018 et 2022.

transition	$0 \rightarrow 0$	$0 \rightarrow 1$	$1 \rightarrow 0$	$1 \rightarrow 1$	$1 \rightarrow 2$	other
proportion	79.62%	7.05%	7.60%	1.15%	0.33%	4.25%

par la procédure récursive suivante :

$$\begin{cases} \alpha_{(i,j)}^n(1) = \rho_i^n \psi_{(i,j)}^n(y_1), \\ \alpha_{(i,j)}^n(t) = \psi_{(i,j)}^n(y_t) p_{i,j}^n \sum_{i' \in \mathbb{N}} \alpha_{(i',i)}^n(t-1), \\ \beta_{(i,j)}^n(T) = 1, \\ \beta_{(i,j)}^n(t-1) = \sum_{j' \in \mathbb{N}} p_{j,j'}^n \psi_{(j,j')}^n(y_t) \beta_{(j,j')}^n(t). \end{cases}$$

Après par exemple n itérations, nous obtenons les paramètres optimaux du modèle et les estimations $\hat{\lambda}, \hat{\mu}, \hat{\nu}$ sont données par

$$(\hat{\lambda}, \hat{\mu}, \hat{\nu}) = g(p_{0,0}^n, p_{0,1}^n, p_{1,0}^n).$$

Il faut noter concernant l'algorithme EM que le résultat dépend fortement de la condition initiale, et il faut également régler le nombre d'itérations et trouver une manière appropriée de tronquer les sommes infinies. Tous ces points sont discutés ainsi que les détails sur l'estimation dans l'article [93]. La Figure 4.1 montre le nombre hebdomadaire de nouveaux cas isolés de fièvre typhoïde entre 2018 et 2022 à Mayotte. Le choix des transitions entre 0 à 1 est motivé par l'observation à la Table 4.3. Quelques exemples de simulation du processus X et Y sont donnés aux Figures 4.2 et 4.3. En terme de perspective, il est prévu l'étude de l'impact de la pluviométrie sur la dynamique de transmission compte tenu des corrélations observées sur des périodes de forte contamination.

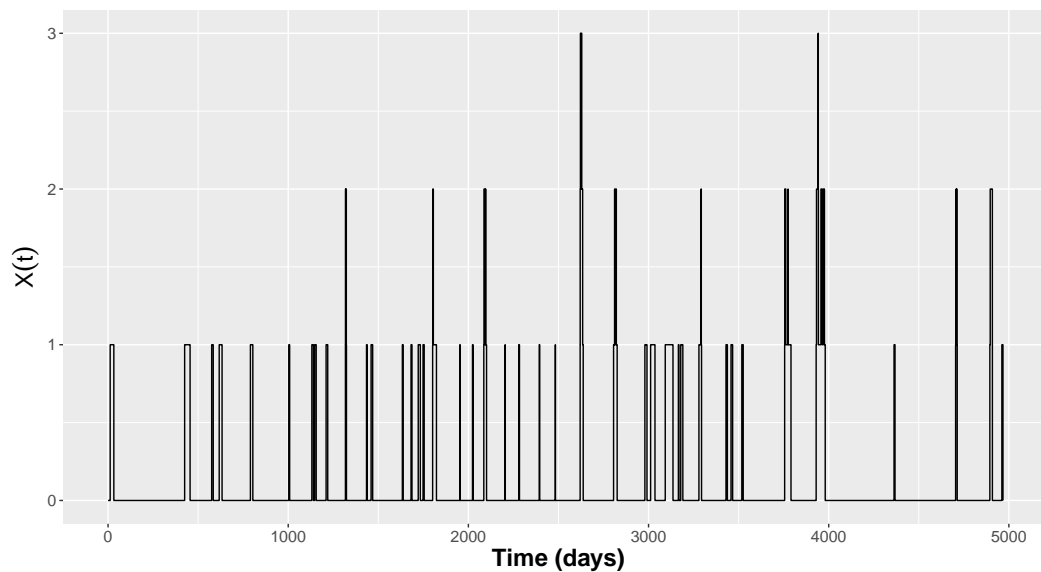


FIGURE 4.2 – Simulation des trajectoires du processus $X(t)$ pour un laps de temps $\Delta t = 1$ jusqu'à un horizon de temps $H = 5000$.

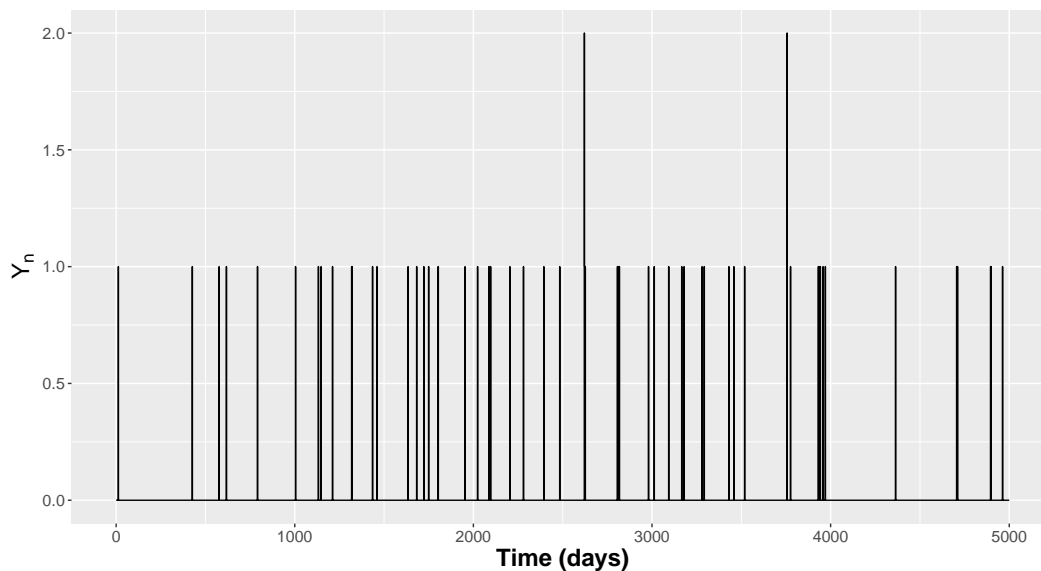


FIGURE 4.3 – Simulation des nouveaux cas infectés Y_n correspondant au processus X pour un laps de temps $\Delta t = 1$ jusqu'à un horizon de temps $H = 5000$.

4.2 Estimation des paramètres d'une solution d'EDS dirigée par un processus stable

En raison de l'augmentation de la puissance de calcul des méthodes statistiques, l'estimation des paramètres pour des équations différentielles stochastiques (EDS) a suscité un grand intérêt au sein de la communauté scientifique. Les méthodes d'estimation classiques et la construction de procédures algorithmiques efficaces et optimales permettent d'entreprendre de nombreuses applications comme la variation des prix en finance à travers les modèles de Cox-Ingersoll-Ross (CIR) et d'Ornstein-Uhlenbeck (OU), ou encore l'étude des dynamiques de population comme les modèles de Lotka Voltera ou des processus de branchement à temps continu. Nous considérons ici un processus stochastique X solution (forte) de l'équation différentielle stochastique suivante :

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dZ_t; \quad X_0 = \eta \quad (4.3)$$

où $Z = \{Z_t, t \geq 0\}$ est un processus de Lévy α -stable ($0 < \alpha < 2$), η est une variable aléatoire indépendante de Z , la fonction b est appelée la dérive et la fonction σ est appelée le coefficient de diffusion. Supposons que ce processus stochastique X est observé à des instants discrets :

$$X_{t_{i+1}} - X_{t_i} = \int_{t_i}^{t_{i+1}} b(X_{s-})ds + \int_{t_i}^{t_{i+1}} \sigma(X_{s-})dZ_s.$$

Nous étudions le problème d'estimation des coefficients. Par exemple, dans le cadre de l'estimation de la fonction de dérive on peut remarquer que :

$$\begin{aligned} \mathbb{E}[X_{i+1} - X_i | X_i = x] &= \mathbb{E}\left[\int_{t_i}^{t_{i+1}} b(X_{s-})ds \middle| X_i = x\right] \\ &+ \mathbb{E}\left[\int_{t_i}^{t_{i+1}} \sigma(X_{s-})dZ_s \middle| X_i = x\right]. \end{aligned}$$

Et en considérant les processus strictement α -stables d'indice $\alpha \in (1, 2)$, l'intégrale stochastique $\int_{t_i}^{t_{i+1}} \sigma(X_{s-})dZ_s$ est une martingale et on a

$$\mathbb{E}[X_{i+1} - X_i | X_i = x] = \mathbb{E}\left[\int_{t_i}^{t_{i+1}} b(X_{s-})ds \middle| X_i = x\right].$$

Puisque $\sup_n |t_{i+1} - t_i| \rightarrow 0$ quand $n \rightarrow \infty$, alors pour de petites variations la fonction de dérive b est supposée constante sur l'intervalle $[t_i, t_{i+1}]$, $n \rightarrow \infty$. Par conséquent

$$\mathbb{E}[X_{i+1} - X_i | X_i = x] \approx b(x)\Delta$$

ou encore

$$b(x) \approx \frac{1}{\Delta} \mathbb{E}[X_{i+1} - X_i | X_i = x]. \quad (4.4)$$

On peut également exprimer la fonction de dérive par polynômes locaux sous des conditions de régularités de la fonction de dérive b . En effet, au voisinage d'une valeur x_0 donnée on a [48] :

$$\begin{aligned} b(x) &= b(x_0) + b'(x_0)(x - x_0) + \frac{b''(x_0)}{2}(x - x_0)^2 + \dots + \frac{b^{(p)}(x_0)}{p!}(x - x_0)^p \\ &= \sum_{j=0}^p \frac{b^{(j)}(x_0)}{j!}(x - x_0)^j = \sum_{j=0}^p \beta_j(x - x_0)^j \end{aligned}$$

avec $\beta_j = \frac{b^{(j)}(x_0)}{j!}$ pour tout j . L'estimation pondérée par polynômes locaux consiste à minimiser sur l'espace des paramètres $\beta_0, \beta_1, \dots, \beta_p$ la fonction de coût suivante :

$$\sum_{i=0}^{n-1} \left(Y_i - \sum_{j=0}^p \beta_j (X_i - x)^j \right)^2 K_h(X_i - x).$$

En particulier pour $p = 1$, la solution est donnée par [48] :

$$\hat{b}(x) = \frac{\sum_{i=0}^{n-1} Y_i \left\{ \frac{S_{n,2}}{h^2} - \left(\frac{X_i - x}{h} \right) \frac{S_{n,1}}{h} \right\} K_h(X_i - x)}{\Delta \sum_{i=0}^{n-1} \left\{ \frac{S_{n,2}}{h^2} - \left(\frac{X_i - x}{h} \right) \frac{S_{n,1}}{h} \right\} K_h(X_i - x)},$$

avec $S_{n,k} = \sum_{i=0}^{n-1} K_h(X_i - x)(X_i - x)^k$, $k = 1, 2$.

On peut également exprimer la fonction de dérive par des méthodes de régression non-paramétriques. Par exemple, l'estimateur non-paramétrique de Nadaraya-Watson proposé par Nadaraya [46] et Watson [47] s'écrit comme suit :

$$\hat{b}_n(x) = \frac{\sum_{i=0}^{n-1} Y_i K_h(X_i - x)}{\Delta \sum_{i=0}^{n-1} K_h(X_i - x)}, \quad \text{où } Y_i = \frac{X_{i+1} - X_i}{\Delta},$$

où $K_h(\cdot) = K(\cdot/h)/h$ et K est un noyau de moyenne nulle et de variance finie. Le paramètre h représente la fenêtre du noyau. Les propriétés de ces estimateurs de la fonction de dérive ont été étudiées dans la littérature. Notre objectif dans cette section est de proposer des estimateurs des paramètres de la partie diffusion en utilisant ces estimateurs non paramétriques de la fonction de dérive. Par exemple, considérons l'EDS dirigée par un processus standard strictement α -stable $(Z_t)_{t \geq 0}$ défini sur l'espace de probabilité filtré $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ sous la forme :

$$\begin{cases} dX_t = f(X_t)dt + \rho g(X_{t-})dZ_t, & t \in [0, T] \\ X_0 = x_0, \end{cases} \quad (4.5)$$

où x_0 est un point de départ du processus, $T > 0$ un horizon de temps donné et la fonction $g : \mathbb{R} \rightarrow \mathbb{R}$ est supposée connue explicitement. Nous supposons que la fonction $f : \mathbb{R} \rightarrow \mathbb{R}$, le paramètre $\rho > 0$ et le processus stable $(Z_t)_{t \text{ dans } [0, T]}$ de paramètres α et β (qui est considéré comme un paramètre de nuisance) sont inconnus. Le noyau $K(\cdot)$ qui est une densité de probabilité symétrique vérifie

$$\sup(1 \vee |u|)K(u) < M_0 < +\infty \quad \text{et} \quad \int_{-\infty}^{+\infty} u^2 K(u)du < +\infty, \quad \int_{-\infty}^{+\infty} K^2(u)du < +\infty.$$

Nous faisons les hypothèses suivantes.

(**A**₁). On suppose que la densité f de la loi invariante π est continue.

(**A**₂). Pour $n \rightarrow \infty$: $h \rightarrow 0$, $\Delta_n \rightarrow 0$,

$$n\Delta_n \rightarrow \infty, \text{ et } \frac{n\Delta_n h}{(\log(n\Delta_n))^2} \rightarrow +\infty.$$

Nous obtenons le résultat suivant.

Théorème 4.4 (Manou-Abi S.M. [87]). *Soit $X = (X_t)_{t \in [0,1]}$ un processus stationnaire réel et fortement mélangeant solution de l'EDS (4.5) et admettant une loi invariante. Nous supposons que le processus X est observé sur des temps discrets t_i avec un pas Δ_n de taille n tel que $g(X_{t_i}) \neq 0$. On suppose que les conditions (\mathbf{A}_1) et (\mathbf{A}_2) sont vérifiées. Soit $(\hat{f}_n(X_{t_i}))_{i=1}^n$ la suite des valeurs estimées de la fonction de dérive par la méthode de Nadaraya-Watson. On définit la fonction caractéristique empirique suivante*

$$\hat{\varphi}_n(u_k) = \frac{1}{n} \sum_{i=0}^{n-1} \exp \left(j u_k \left(\frac{\Delta X_{t_i} - \hat{f}_n(X_{t_i}) \Delta_n}{g(X_{t_i})} \right) \right) \quad (4.6)$$

pour tout $k \in [[1, m]]$, $m > 0$, et $j^2 = -1$. Posons

$$\left\{ \begin{array}{l} \hat{\alpha}_m = \frac{\sum_{k=1}^m W_k V_k - \frac{1}{m} \sum_{k=1}^m V_k \sum_{k=1}^m W_k}{\sum_{k=1}^m W_k^2 - \frac{1}{m} \left(\sum_{k=1}^m W_k \right)^2}, \\ \hat{\lambda}_m = \frac{1}{m} \sum_{k=1}^m V_k - \frac{\hat{\alpha}_m}{m} \sum_{k=1}^m W_k \\ \hat{\rho}_m = \Delta_n^{-\frac{1}{\hat{\alpha}_m}} \exp \frac{\hat{\lambda}_m}{\hat{\alpha}_m} \\ \hat{\beta}_m = \frac{\sum_{k=1}^m u_k S_k}{\sum_{k=1}^m u_k B_k}, \end{array} \right. \quad (4.7)$$

où

$$S_k = \arg(\hat{\varphi}_n(u_k)), \quad B_k = \tan\left(\frac{\pi \hat{\alpha}}{2}\right) \text{sign}(u_i) (u_k - u_k^{\hat{\alpha}_m}) \Delta_n^{1/\hat{\alpha}_m}, \\ V_k = \log(-\log|\hat{\varphi}_n(u_k)|) \quad \text{et} \quad W_k = \log(|u_k|).$$

Alors $\hat{\alpha}_m$, $\hat{\rho}_m$ et $\hat{\beta}_m$ sont des estimateurs des moindres carrés de α , ρ et β .

En fonction de la normalité des distributions observées des estimées de $\hat{\varphi}_n(u_k)$, ces estimateurs peuvent être consistents et robustes. La valeur optimale de m est discutée dans [?], préconisant la sélection de points u_k dans l'intervalle $[0.1, 1]$. Il faut noter qu'il existe des transformations de normalisation des données. Le Min-Max Scaling peut être appliqué quand les données varient dans des échelles différentes et permet de réduire l'effet des outliers. Les conditions de mélange sont liées aux propriétés d'ergodicité de l'EDS 4.5 et sont détaillées dans l'article [87] pour évaluer la performance des estimateurs sur données simulées. Ce résultat est nouveau, eu égard à la littérature existante sur les estimateurs des solutions d'EDS dirigées par des processus α -stables. Notre méthodologie est basée sur la méthode d'estimation par la fonction caractéristique décrite ci-dessus et la méthode des moindres carrés (linéaire ou pondérée). Nous discutons de la validité et de l'efficacité de l'implémentation numérique des estimateurs sur des données générées à partir d'une solution vérifiant les hypothèses du Théorème dans les tableaux 4.4 et 4.5. Dans le résultat qui suit, nous discutons de l'estimation paramétrique des coefficients de dérive θ et λ des modèles CIR et OU.

Théorème 4.5 (Manou-Abi S.M. [87]). *Soit $X = (X_t)_{t \in [0,1]}$ un processus stationnaire réel et fortement mélangeant solution de l'EDS (3.5) ou (3.6). Nous supposons que le*

processus X est observé sur des temps discrets t_i avec un pas Δ fixé. On suppose que les conditions (\mathbf{A}_2) et (\mathbf{A}_3) sont vérifiées. Soit $(\hat{f}_n(X_{t_i}))_{i=1}^n$ la suite des valeurs estimées de la fonction de dérive par la méthode de Nadaraya-Watson. Posons

$$\left\{ \begin{array}{l} \hat{\alpha}_n = \frac{\sum_{i=0}^{n-1} X_{t_i} \hat{f}_n(X_{t_{i+1}}) - \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i} \sum_{i=1}^n \hat{f}_n(X_{t_{i+1}})}{\frac{1}{n} \left(\sum_{i=0}^{n-1} X_{t_i} \right)^2 - \sum_{i=0}^{n-1} X_{t_i}^2}, \\ \hat{\lambda}_n = \frac{1}{n} \sum_{i=0}^{n-1} \hat{f}_n(X_{t_{i+1}}) - \hat{\alpha}_n \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i}, \\ \hat{\theta}_n = -\frac{1}{\Delta} W(-\hat{\alpha}_n \Delta), \end{array} \right. \quad (4.8)$$

où W est la fonction Lambert, voir [57].

— Nous supposons de plus que $q = \alpha^{-1}$ avec $\alpha \in (\sqrt{2}, 2)$ et le processus α -stable est de saut positifs dans (3.6).

Pour n suffisamment grand, $\hat{\theta}_n$ et $\hat{\lambda}_n$ sont des estimateurs consistants et sans biais de θ et λ .

Des applications à des données réelles de variation des prix d'achat sont également fournies dans [87]. En terme de perspective, je vais non seulement étudier la loi asymptotique et l'erreur mais aussi mettre en place un package R pour ce type de problème d'estimation.

TABLE 4.4 – Performance des estimateurs des paramètres $(\hat{\alpha}, \hat{\beta}, \hat{\rho})$ dans le cadre d'un modèle d'OU dirigé par un processus standard 1.6-stable symétrique

n et T	Vrais paramètres	Paramètres estimés	RMSE
500 et T=1	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.70, \quad \hat{\beta}_n = 0.08, \quad \hat{\rho}_n = 0.80$	1.55
500 et T=10	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.69, \quad \hat{\beta}_n = 0.015, \quad \hat{\rho}_n = 3.38$	8.34
1000 et T=1	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.55, \quad \hat{\beta}_n = 0.11, \quad \hat{\rho}_n = 1.2$	1.10
1000 et T=10	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.56, \quad \hat{\beta}_n = 0.13, \quad \hat{\rho}_n = 4.87$	8.43
2000 et T=1	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.6346, \quad \hat{\beta}_n = 0.1146, \quad \hat{\rho}_n = 0.9335$	0.77
2000 et T=10	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.6029, \quad \hat{\beta}_n = 0.064, \quad \hat{\rho}_n = 4.33$	6.91

TABLE 4.5 – Performance des estimateurs des paramètres $(\hat{\alpha}, \hat{\beta}, \hat{\rho})$ dans le cadre d'un modèle CIR dirigé par un processus standard symétrique 1.6-stable

n et T	Vrais paramètres	Paramètres estimés	RMSE
500 et T=1	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.66, \quad \hat{\beta}_n = 0.13, \quad \hat{\rho}_n = 0.88$	1.17
500 et T=10	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.51, \quad \hat{\beta}_n = 0.08, \quad \hat{\rho}_n = 5.95$	8.24
1000 et T=1	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.63, \quad \hat{\beta}_n = 0.217, \quad \hat{\rho}_n = 0.98$	3.72
1000 et T=10	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.53, \quad \hat{\beta}_n = 0.35, \quad \hat{\rho}_n = 6.32$	15.36
2000 et T=1	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.61, \quad \hat{\beta}_n = 0.07, \quad \hat{\rho}_n = 1.03$	2.07
2000 et T=10	$\alpha = 1.6, \quad \beta = 0, \quad \rho = 1$	$\hat{\alpha}_n = 1.64, \quad \hat{\beta}_n = 0.04, \quad \hat{\rho}_n = 4$	11.13

Statistique computationnelle, statistique des extrêmes

L'objectif de ce chapitre est de présenter ma contribution relative d'une part à l'estimation des paramètres des densités de probabilités de certaines familles de lois, y compris les lois α -stables, et d'autre part à l'estimation de l'indice de queue des lois à queues lourdes, par la théorie des valeurs extrêmes. Les publications relatives à ces travaux se retrouvent dans [69, 70, 71, 85] et [94].

5.1 Lois stables et estimation non paramétrique du taux de reproduction effectif

Je m'intéresse dans cette partie à l'estimation statistique du taux de reproduction effectif dans la transmission de l'infection liée au Covid-19 par l'approche non-paramétrique suivante :

$$R_0(t) = \frac{\Gamma(t)}{\sum_{\tau \leq t} d(\tau)\Gamma(t - \tau)},$$

où $\Gamma(t)$ est le nombre de nouveaux infectés au temps t et d la densité de distribution de l'intervalle sériel mesurant l'intervalle de temps entre l'apparition des symptômes d'un infecteur donné et l'apparition des symptômes chez l'infecté. Ce dernier est essentiellement influencé par la période d'incubation de l'infecté (temps passé entre le moment où l'infecté est infecté par son infecteur et le moment où il pourrait développer des symptômes et être contagieux) et les différents profils d'infectiosité (l'intervalle de temps depuis l'apparition des symptômes d'un infecté à la contamination à une autre personne). Dans un premier temps, je me suis intéressé à l'estimation de l'intervalle sériel via une base de données collectée chez des patients suivis et confirmés en laboratoire [67] par des méthodes d'estimation de la log-vraisemblance par mélange de lois. L'estimation par mélange est un outil flexible pour modéliser une fonction de densité de probabilité comme une somme pondérée de fonctions de densité dans des observations indépendantes et provenant de sources multivariées. Le taux de reproduction est un indicateur pour évaluer le nombre moyen de nouveaux cas d'infection, engendré par un individu au cours d'une période d'infectiosité. Il dépend principalement des facteurs précédemment évoqués : durée de la contagiosité après infection, probabilité d'obtenir une infection après un contact entre une personne

infectée et une personne susceptible et aussi de la fréquence des contacts humains. En considérant des intervalles sériels (IS) positifs, et des modèles de mélange basés sur des distributions de la famille des lois d'extrêmes généralisées (Gumbel, Fréchet, Weibull), nous obtenons les résultats suivants [71] avec les données Covid-19 de l'île de Mayotte (Figure 5.1). Ces résultats viennent améliorer les méthodes existantes d'estimation du taux de reproduction effectif disponible dans la littérature. Considérons maintenant le modèle mathématique donné par le système d'équations différentielles suivant :

$$\begin{cases} dS(t) = -\beta_p S \frac{(I_s + I_a)}{N} dt + \omega R dt \\ dE(t) = \beta_p S \frac{(I_s + I_a)}{N} dt - \lambda E dt \\ dI_s(t) = (1 - q)\lambda E dt - (\gamma_1 + \alpha) I_s dt \\ dI_a(t) = q\lambda E dt - \gamma_2 I_a dt \\ dR(t) = \gamma_1 I_s dt + \gamma_2 I_a dt - \omega R dt \\ dD(t) = \alpha I_s dt \end{cases}$$

avec $N = S + E + I_s + I_a + R + D$ la taille (fixée) de la population totale et la population répartie suivant les compartiments de Susceptibles **S**, Exposés **E** (infecté mais pas encore infectieux) dans une période homogène de latence λ^{-1} ; Infectés symptomatiques **I_s**, avec une période infectieuse γ_1^{-1} ; Infectés asymptomatiques **I_a** dans une période infectieuse γ_2^{-1} ; Guéris **R**, et cas de morts **D**. Nous supposons qu'une fraction ω des individus guéris redeviennent susceptibles (perte d'immunité et réinfection). La méthode basée sur la matrice dite "Next-generation" de Dieckmann permet d'obtenir

$$R_0 = \frac{q\beta_p}{\gamma_2} + \frac{(1-q)\beta_p}{\gamma_1 + \alpha} \text{ et } \beta_p \sim \frac{R_0\gamma_1}{r + 1 - q}, \quad r := \frac{\gamma_1}{\gamma_2},$$

en négligeant le taux de mortalité. On obtient des estimations du paramètre de transmission β dans la Figure 5.2. Une extension de ce travail a été faite pour prendre en compte les distributions négatives de l'intervalle sériel en introduisant les lois α -stables. Les lois α -stables non-gaussiennes sont un sous-ensemble de lois à queue régulière. Malheureusement, il n'existe pas d'expression simple de la densité de probabilité à l'exception des distributions gaussiennes, de Lévy $\alpha = 1/2$ et de Cauchy $\alpha = 1$. Seuls les moments d'ordre $p < \alpha$ existent et dans le cas $\alpha \in (0, 1]$ les espérances explosent. L'estimation des paramètres est un problème encore ouvert et riche en littérature. Ainsi dans [90], nous avons considéré l'estimation de la distribution de l'intervalle sériel en prenant en compte des mélanges de distributions α -stables univariées et l'estimation des paramètres associés. Nous avons proposé et adapté plusieurs méthodes basées sur les fonctions caractéristiques, les estimateurs à noyau gaussien des distribution de densité, l'algorithme EM et une méthode d'estimation bayésienne. La performance des méthodes qui sont ensuite appliquées à l'estimation du taux de reproduction effectif, semblent meilleures avec les distributions α -stables comparé aux travaux précédents. Quelques résumés des méthodes considérées.

5.1.1 Estimation de paramètres de lois stables par fonctions caractéristique et score

Supposons que la loi de distribution de l'intervalle sériel suit une loi stable $X \sim S(\alpha, \beta, \gamma, \zeta; 0)$ et soit X_1, \dots, X_n un échantillon observé de taille n de X , de fonction

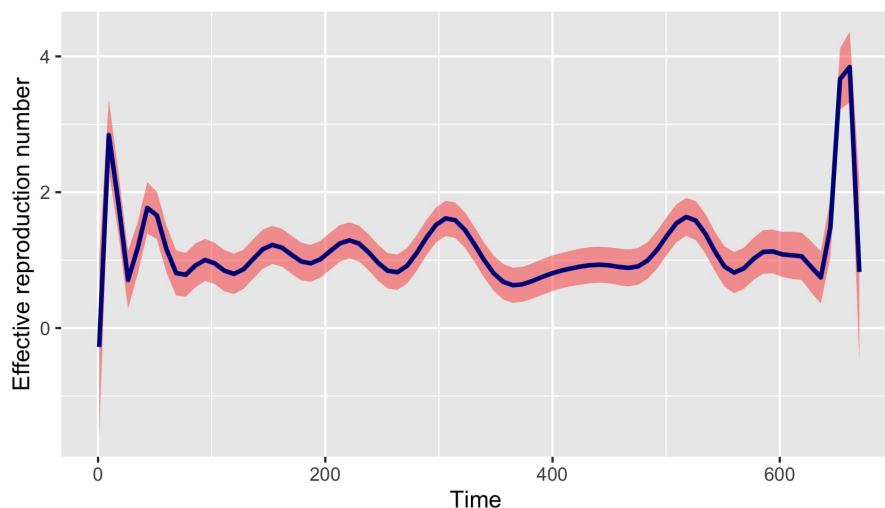


FIGURE 5.1 – Estimation du taux de reproduction à Mayotte du 13 mars 2020 au 11 janvier 2022, avec le meilleur modèle de mélange de lois ajusté pour l’IS positif.

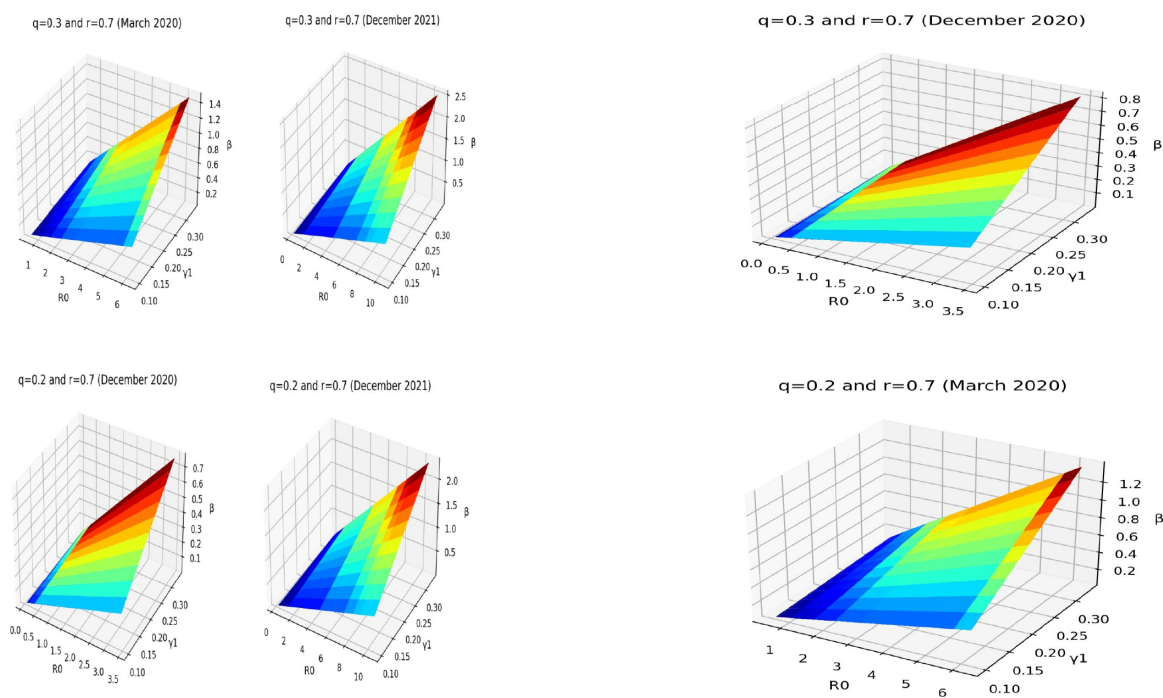


FIGURE 5.2 – Estimations du taux de transmission en fonction d’une plage de périodes infectieuses γ_1^{-1} dans $(3, 10)$.

caractéristique empirique donnée par

$$\hat{\phi}_n(t) = \int_{\mathbf{R}} \exp(jtx) dF_n(x) = \frac{1}{n} \sum_{k=1}^n \exp(jtX_k).$$

Nous introduisons un estimateur alternatif de la fonction caractéristique en utilisant le noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$ pour tout $x \in \mathbf{R}$ avec $\int_{\mathbf{R}} K(z) dz = 1$. On obtient ainsi l'estimateur suivant de la fonction caractéristique :

$$\hat{\phi}_n(t) = \frac{1}{nh_n} \int_{\mathbf{R}} \exp(itx) \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right) dx,$$

où la fenêtre $h_n \rightarrow 0$. Pour des raisons pratiques et dans le cas $\alpha > 1$, un choix optimal de h_n peut se faire en considérant la méthode proposée par Sheather et Jones [52] ou de manière générale la méthode d'intégration de deuxième génération proposée par Slaoui dans [53]. En prenant le logarithme de $\hat{\phi}_n$ on obtient les relations linéaires suivantes : $y_k = +\alpha \log |u_k| + \log(\gamma^\alpha)$ et $z_k = \beta B_k + \zeta u_k$, avec

$$\begin{cases} y_k &= \log(-\log(|\hat{\phi}_n(t_k)|)), \\ z_k &= \arg(\hat{\phi}_n(t_k)), \\ B_k &= \hat{\gamma}^{\hat{\alpha}} \eta(\hat{\gamma} t_k | \hat{\alpha}; 0). \end{cases}$$

On obtient alors les estimateurs suivants par la méthode des moindres carrés linéaire [90].

$$\begin{aligned} \hat{\alpha}_m &= \left(\sum_{k=1}^m W_k \log(|t_k|) y_k - \frac{\sum_{k=1}^m W_k y_k}{\sum_{k=1}^m W_k} \times \sum_{k=1}^m W_k \log(|t_k|) \right) \\ &\quad \times \left(\sum_{k=1}^m W_k \log(|t_k|)^2 - \frac{\sum_{k=1}^m W_k \log(|t_k|)}{\sum_{k=1}^m W_k} \times \sum_{k=1}^m W_k \log(|t_k|) \right)^{-1}, \\ \hat{\alpha}_m &= \frac{\sum_{k=1}^m W_k y_k - \hat{\alpha}_m \sum_{k=1}^m W_k \log(|t_k|)}{\sum_{k=1}^m W_k}, \\ \hat{\gamma}_m &= \exp\left(\frac{\hat{\alpha}_m}{\hat{\alpha}_m}\right). \end{aligned}$$

et

$$\begin{aligned} \hat{\zeta}_m &= \left(\sum_{k=1}^m W_k B_k z_k - \frac{\sum_{k=1}^m W_k t_k z_k}{\sum_{k=1}^m W_k t_k B_k} \times \sum_{k=1}^m W_k B_k^2 \right) \\ &\quad \times \left(\sum_{k=1}^m W_k B_k t_k - \frac{\sum_{k=1}^m W_k t_k^2}{\sum_{k=1}^m W_k t_k B_k} \times \sum_{k=1}^m W_k B_k^2 \right)^{-1}, \\ \hat{\beta}_m &= \frac{\sum_{k=1}^m W_k t_k z_k - \hat{\zeta}_m \sum_{k=1}^m W_k t_k^2}{\sum_{k=1}^m W_k t_k B_k}. \end{aligned}$$

où $W = (W_k)_k$ désigne une suite de poids donnée aux observations. Le choix optimal de m et des points t_k a été discuté dans la littérature [54]. Ces estimateurs peuvent être non biaisés et consistents en fonction de la normalité derrière la distribution des valeurs de y_k générées. Il faut noter toutefois qu'il existe de nombreuses transformations pour ramener une série de données à des distributions normales.

Maintenant, dans le cadre de méthodes d'approximation numérique basées sur des fonctions scores (dérivée partielle de la log-vraisemblance par rapport au paramètre inconnu), posons pour $d > 1$ et $\eta(r, \alpha; 1) = -\tan(\frac{\pi\alpha}{2}r^\alpha)$:

$$\begin{aligned} g_d(x|\alpha, \beta) &= \int_0^\infty \cos(xr + \beta\eta(r, \alpha; 1))r^{d-1} \exp(-r^\alpha)dr, \\ \tilde{g}_d(x|\alpha, \beta) &= \int_0^\infty \sin(xr + \beta\eta(r, \alpha; 1))r^{d-1} \exp(-r^\alpha)dr, \\ h_d(x|\alpha, \beta) &= \int_0^\infty \cos(xr + \beta\eta(r, \alpha; 1)) \log(r)r^{d-1} \exp(-r^\alpha)dr \\ \tilde{h}_d(x|\alpha, \beta) &= \int_0^\infty \sin(xr + \beta\eta(r, \alpha; 1)) \log(r)r^{d-1} \exp(-r^\alpha)dr. \end{aligned}$$

Le résultat suivant contenu dans [34] permet d'évaluer les fonctions scores correspondant aux paramètres de la loi α -stable.

Théorème 5.1 (Fonction score [34]). *Soit $\alpha \neq 1$. Les fonctions de score sur chaque paramètres de la loi stable sont données par :*

$$\begin{aligned} \frac{\partial f}{\partial \alpha}(x|\alpha, \beta, \gamma, \zeta; 1) &= \frac{1}{\pi\gamma} \left[\frac{\pi\beta}{2 \cos(\frac{\pi\alpha}{2})^2} \tilde{g}_{1+\alpha} \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right) \right. \\ &\quad \left. + \beta \tan(\frac{\pi\alpha}{2}) \tilde{h}_{1+\alpha} \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right) - h_{1+\alpha} \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right) \right], \\ \frac{\partial f}{\partial \beta}(x|\alpha, \beta, \gamma, \zeta; 1) &= \frac{\tan(\frac{\pi\alpha}{2})}{\pi\gamma} \tilde{g}_{1+\alpha} \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right), \\ \frac{\partial f}{\partial \gamma}(x|\alpha, \beta, \gamma, \zeta; 1) &= -\frac{1}{\pi\gamma^2} g_1 \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right) + \frac{x-\zeta}{\pi\gamma^3} \tilde{g}_2 \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right), \\ \frac{\partial f}{\partial \zeta}(x|\alpha, \beta, \gamma, \zeta; 1) &= \frac{1}{\pi\gamma^2} \tilde{g}_2 \left(\frac{x-\zeta}{\gamma} | \alpha, \beta \right). \end{aligned}$$

Nous voyons ainsi qu'on peut faire appel aux algorithmes d'implémentation numérique des zéros d'une fonction pour estimer les paramètres concernés au travers de dérivées des fonctions de log-vraisemblance (scores).

5.1.2 Estimation de paramètres de lois stables par les algorithmes stochastiques

Comme évoqué plus haut les méthodes d'estimation par mélange de distributions sont des outils pratiques et efficaces. Deux méthodes d'approximation stochastiques courantes d'inférence des paramètres dans les modèles de mélange sont : l'algorithme EM et l'approche bayésienne [55, 56]. L'algorithme EM est une méthode utilisée pour estimer les paramètres des modèles statistiques avec des variables latentes ou manquantes. Cet algorithme est particulièrement utile dans les cas où les données sont incomplètes ou partiellement observées. L'estimation bayésienne est un cadre pour la formulation de problèmes d'inférence statistique. Dans la prédiction ou l'estimation d'une variable aléatoire la méthode bayésienne est basée sur la connaissance préalable de la distribution de probabilité de la variable aléatoire. La méthode d'estimation bayésienne utilise des données préalables pour estimer la valeur des paramètres inconnus. Cela réduit la différence

entre l'estimateur et la valeur réelle de ce paramètre. Dans la modélisation bayésienne, la sélection des lois à priori joue donc un rôle crucial dans l'inférence à posteriori. Nous présentons ici une version adaptée de l'algorithme EM en incluant aussi l'estimation par la méthode de la fonction caractéristique (ECF). On désigne toujours par n le nombre d'observations et z_i les observations latentes pour $i = 1, \dots, n$.

Algorithm 2 Algorithme EM pour un mélange de lois α -stable (Hajjaji O., Manou-Abi S.M. et Slaoui Y. [90])

- 1: Initialisation des paramètres et choix du seuil d'erreur ϵ .
- 2: **repeat**
- 3: **Etape-E :**
- 4: Evaluation de la probabilité a posteriori
- 5: **for** $i = 1, \dots, n$ **do**
- 6: **for** $j = 1, 2$ **do**
- 7: L'observation i appartient à la composante j ($z_i = j$) avec probabilité

$$p_{i,j}^{(t)} = \frac{\lambda^{(t)} f_j(x_i | \alpha_1^{(t)}, \beta_1^{(t)}, \gamma_1^{(t)}, \zeta_1^{(t)}; 0)}{\lambda^{(t)} f_1(x_i | \alpha_1^{(t)}, \beta_1^{(t)}, \gamma_1^{(t)}, \zeta_1^{(t)}; 0) + (1 - \lambda^{(t)}) f_2(x_i | \alpha_2^{(t)}, \beta_2^{(t)}, \gamma_2^{(t)}, \zeta_2^{(t)}; 0)}$$

- 8: **end for**
 - 9: **end for**
 - 10: **Etape-M :**
 - 11: **for** $j = 1, 2$ **do**
 - 12: $\lambda_j^{(t+1)} = \frac{1}{n} \{\#z_i = j\}$.
 - 13: Ensuite, nous utilisons la méthode d'estimation par log-vraisemblance approchée ou fonction caractéristique pour obtenir $\Theta_j^{(t+1)} = (\alpha_j^{(t+1)}, \beta_j^{(t+1)}, \gamma_j^{(t+1)}, \zeta_j^{(t+1)})$
 - 14: Nous calculons la log-vraisemblance à l'itération $t + 1$, notée $Q^{(t+1)}$.
 - 15: **end for**
 - 16: **until** $|Q^{(t+1)} - Q^{(t)}| < \epsilon$.
-

Concernant l'approche bayésienne, un point important est la connaissance préalable de la distribution dont les paramètres sont issus. Ces connaissances préalables proviennent généralement de l'expérience ou d'expériences antérieures. Ainsi on suppose par exemple que les prioris des paramètres α et β sont de lois uniformes, un prior conjugué Inverse Gamma (IG) pour le paramètre d'échelle γ et un prior de loi normale pour le paramètre de position ζ . De plus nous supposons que les poids sont distribués suivant une loi de Dirichlet symétrique et de manière indépendante des autres distributions. Il faut noter que le calcul de la distribution postérieure conjointe des paramètres susmentionnés est souvent difficile à analyser pour les densités α -stables compte tenu de l'absence de formules explicites et analytiques. Pour surmonter ce problème, on peut utiliser des méthodes de Monte Carlo par chaîne de Markov (MCMC), plus les algorithmes d'échantillonnage de Gibbs ou de Metropolis-Hastings, comme illustré dans les étapes suivantes.

Loi de distribution à priori des poids

La loi de Dirichlet (\mathcal{D}) est une loi bien connue et utile pour les données de proportion. Nous supposons que la distribution a priori du poids $\lambda = (\lambda_1, \lambda_2)$ suit une loi de Dirichlet (ici Beta) symétrique de paramètres $\xi = (1, 1)$. Puisque $\mathbb{P}(z_i = j)$ est égal à λ_j pour $j = 1, 2$ et $i = 1, \dots, N$, où N est le nombre d'observations; alors la distribution

conditionnelle complète pour λ suit également une loi de Dirichlet (ici Beta), avec des paramètres $\xi_j + n_j$, où n_j est la fréquence d'observations affectées à la composante j . Ainsi, la distribution pour les poids est $\lambda|\Theta \sim \mathcal{D}(\xi_1 + n_1, \xi_2 + n_2)$.

Mise à jour des paramètres par MCMC

Nous considérons la méthode d'échantillonnage de Metropolis-Hastings. Nous générons un paramètre candidat $\Theta_j^{new} = (\alpha_j^{new}, \beta_j^{new}, \gamma_j^{new}, \zeta_j^{new})$, à partir d'une distribution de loi $q(\cdot|\cdot)$ acceptée avec probabilité $A_{\Theta_j^{new}}$, définie par : $A_{\Theta_j^{new}} =$

$$\min \left(1, \prod_{i=1, z_i=j}^N \frac{f(x_i|\alpha_j^{new}, \beta_j^{new}, \gamma_j^{new}, \zeta_j^{new}; 0)}{f(x_i|\alpha_j^{old}, \beta_j^{old}, \gamma_j^{old}, \zeta_j^{old}; 0)} \times \frac{p(\Theta_j^{new})q(\Theta_j^{old}|\Theta_j^{new})}{p(\Theta_j^{old})q(\Theta_j^{new}|\Theta_j^{old})} \right).$$

Etant donné que les lois à priori sont supposées indépendantes, alors on obtient :

$$p(\Theta) = p(\alpha)p(\beta)p(\gamma)p(\zeta).$$

En choisissant une loi normale pour $q(\cdot|\cdot)$, on obtient par symétrie :

$$q(\Theta_j^{new}|\Theta_j^{old}) = q(\Theta_j^{old}|\Theta_j^{new}).$$

Ainsi $A_{\Theta_j^{new}}$ devient

$$\min \left(1, \prod_{i=1, z_i=j}^N \frac{f(x_i|\alpha_j^{new}, \beta_j^{new}, \gamma_j^{new}, \zeta_j^{new}; 0)}{f(x_i|\alpha_j^{old}, \beta_j^{old}, \gamma_j^{old}, \zeta_j^{old}; 0)} \times \frac{IG(\gamma_j^{new}|\alpha_0, \beta_0)N(\omega_j^{new}|\epsilon, k)}{IG(\gamma_j^{old}|\alpha_0, \beta_0)N(\zeta_j^{old}|\epsilon, k)} \right). \quad (5.1)$$

On génère une loi uniforme u dans $[0, 1]$ de sorte que si $A_{\Theta_j^{new}} > u$, on accepte les nouvelles valeurs, sinon on conserve celles de l'itération précédente. Le fait que nous considérons une seule zone de rejet associée au paramètre vectoriel Θ est possible puisque les lois à priori sont supposées indépendantes. Ainsi, la chaîne de Markov $\tilde{\Theta}_n = (\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\gamma}_n, \tilde{\zeta}_n)$ est stationnaire (où n est l'indice d'itération). Cette façon de faire semble être numériquement plus précise que de considérer différentes zones de rejet pour chaque paramètre séparément, contrairement à [56] où les auteurs considèrent une multi chaîne de Markov pour chaque paramètre sans tirer profit de l'indépendance.

Mis à jour des poids

Il est nécessaire, à chaque itération, de déterminer à quelle sous-population j appartient chaque observation prédicte. Pour cela nous évaluons la probabilité que l'observation x_i appartienne à la composante j :

$$\mathbb{P}(z_i = j|\Theta) = \frac{\lambda_j f_j(x_i|\alpha_j, \beta_j, \gamma_j, \zeta_j; 0)}{\lambda_1 f_1(x_i|\alpha_1, \beta_1, \gamma_1, \zeta_1; 0) + \lambda_2 f_2(x_i|\alpha_2, \beta_2, \gamma_2, \zeta_2; 0)}. \quad (5.2)$$

Nous présentons sur les lignes suivantes une version adaptée de l'algorithme par cette approche bayésienne.

Algorithm 3 Approche bayésienne pour un mélange de distributions α -stables (Hajjaji O., Manou-Abi S.M. et Slaoui Y. [90])

Input: Initialisation des paramètres $\Theta = (\alpha, \beta, \gamma, \zeta)$

Input: Considérons un nombre d'itérations N .

- 1: **for** $t = 1, \dots, N$ **do**
 - 2: Générer des poids $\lambda = (\lambda_1, \lambda_2)$ suivant une loi de Dirichlet (Beta car deux composantes) symétrique $\lambda \sim D(\xi_1 + n_1, \xi_2 + n_2)$ avec n_1 la fréquence d'observations de la première composante et n_2 celle de la seconde composante.
 - 3: Implémenter la loi $q(\cdot|\cdot) = N(\cdot|\theta, \sigma)$, où θ désigne les valeurs précédentes du paramètre Θ et σ un écart type donné assez petit.
 - 4: Générer les nouvelles valeurs $\Theta_j^{new} = (\alpha_j^{new}, \beta_j^{new}, \gamma_j^{new}, \zeta_j^{new})$ à partir de la loi $q(\cdot|\cdot) = N(\cdot|\theta, \gamma)$ pour chaque composante.
 - 5: Acceptez Θ_j^{new} suivant l'équation (5.1) et posez $\Theta_j^t = \Theta_j^{new}$, sinon prendre $\Theta_j^t = \Theta_j^{t-1}$.
 - 6: **for** chaque observation x_i **do**
 - 7: Obtenir la variable de position z_i à l'aide de l'équation (5.2).
 - 8: **end for**
 - 9: **end for**
 - 10: Calculer les paramètres moyens jusqu'à une période donnée M notée **burn-in** : $\Theta_j = \frac{1}{N-M} \sum_{k=M}^N \Theta_j^{(k)}$.
-

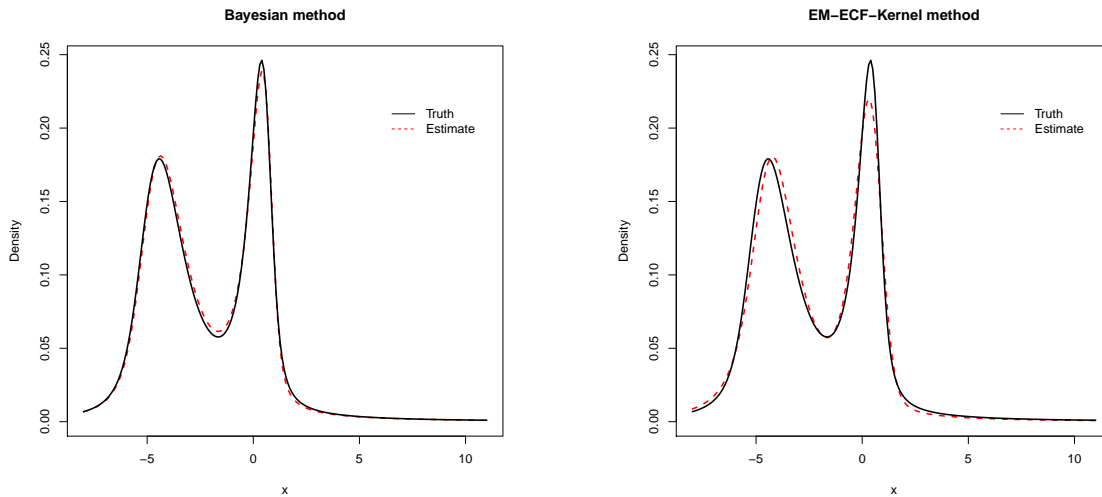
Des exemples numériques illustrant la performance des méthodes de la fonction caractéristique (ECF-Kernel), du Maximum de vraisemblance par fonction de score (ML-method), de l'algorithme EM et de la méthode bayésienne sont présentées dans les Tableaux 5.1 et 5.2. Une application aux données d'intervalles sériels et Covid-19 à l'île de Mayotte mentionnées plus haut concernant la variation du taux de reproduction observé est donné à la Figure 5.3.

TABLE 5.1 – Comparaison EM-Configuration 1

Paramètre	Vrai valeur	n	ECF-Kernel	ML-method
α	1.6	500	1.6356	1.6023
		750	1.6191	1.5484
		1000	1.6075	1.5519
β	-0.8	500	0.7541	-0.7566
		750	-0.8375	-0.7668
		1000	-0.7696	-0.7337
γ	5	500	5.0124	4.8583
		750	5.2009	4.9608
		1000	5.1666	4.9888
ζ	12	500	11.8914	11.9563
		750	12.2187	12.2930
		1000	12.1193	12.1905

TABLE 5.2 – Comparaison EM-Configuration 1

Paramètre	Vrai valeur	n	ECF-Kernel	ML-method
α	1.4	500	1.3848	1.3243
		750	1.3973	1.3613
		1000	1.3329	1.2985
β	0.5	500	0.4304	0.4464
		750	0.5173	0.5651
		1000	0.5381	0.5542
γ	2	500	1.9819	1.8854
		750	2.1368	2.0573
		1000	2.0910	2.0162
ζ	-10	500	-9.9746	-10.0041
		750	-9.9876	-10.0300
		1000	-10.0807	-10.1048



5.2 Estimation de lois à queues lourdes par la théorie des valeurs extrêmes

La théorie des valeurs extrêmes (Extreme Value Theory, EVT en anglais), est une branche importante de la théorie des probabilités et de la statistique qui permet de modéliser le comportement des événements rares et/ou extrêmes. La distribution limite des valeurs extrêmes est indexée par un paramètre appelé l'indice de queue. La connaissance de cet indice est important pour la modélisation. Soit $(X_n)_{n \geq 1}$ une suite de copies indépendantes d'une variable aléatoire X de fonction de répartition $F(x) = \mathbb{P}(X \leq x)$. On note $X_{1,n} \leq \dots \leq X_{n,n}$ l'échantillon ordonné, pour tout $i = 1, \dots, n$ la variable aléatoire $X_{i,n}$ s'appelle la i ème statistique d'ordre n de l'échantillon. La théorie des valeurs extrêmes se focalise sur l'étude statistique de la tendance de $M_n = \max(X_1, \dots, X_n)$. Plus précisément, on étudie le comportement limite de la quantité $\frac{M_n - b_n}{a_n}$ où $a_n > 0$ et $b_n \in \mathbb{R}$ sont des constantes de normalisation pour tout $n \geq 1$. Le Théorème de Fisher-Tippett-Gnedenko (un des fondements de la théorie des valeurs extrêmes) considère que s'il existe de telles constantes de normalisation pour lesquelles la suite $(\frac{M_n - b_n}{a_n})_{n \geq 1}$ converge en loi vers une distribution G non dégénérée, alors G appartient à la famille des lois de valeurs

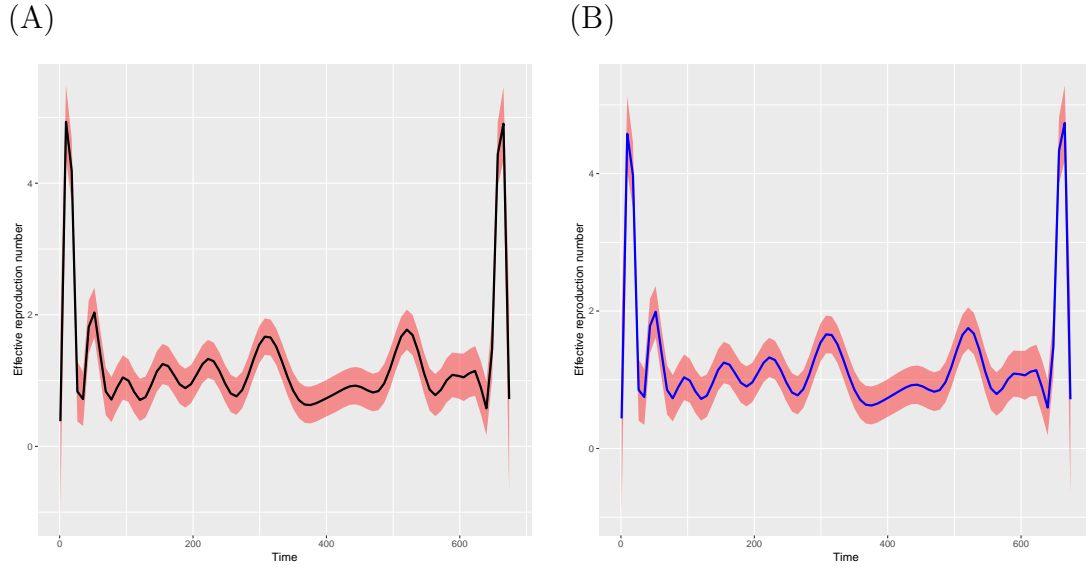


FIGURE 5.3 – Variation du taux de reproduction à Mayotte du 13 mars 2020 au 11 janvier 2022 avec un mélange de lois α -stable sur l’intervalle sériel par l’approche Bayésienne (A), et l’approche EM incluant la méthode ECF à noyaux (B).

extrêmes généralisée (Generalized Extreme Value) notée GEV.

Les notations des paramètres dans cette section ne doivent pas être confondues avec celles des sections précédentes.

Définition 7 ([91]). *La densité de probabilité d’une distribution GEV de paramètre de localisation $\mu \in \mathbb{R}$, d’échelle ou de dispersion $\sigma > 0$ et de forme $\gamma \in \mathbb{R}$ est*

$$f(z/\mu, \sigma, \varepsilon) = \begin{cases} \exp\left(-\left(1 + \gamma\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\gamma}\right) & \text{si } \gamma \neq 0, \\ \exp\left(-\left(\frac{z-\mu}{\sigma}\right)^{-1/\gamma}\right) & \text{si } \gamma = 0 \end{cases}$$

pour tout z telle que $1 + \gamma\left(\frac{z-\mu}{\sigma}\right) > 0$.

Le paramètre γ est appelé également l’indice des valeurs extrêmes. Selon le signe de ce paramètre, on distingue trois familles de lois appelées domaines d’attraction. Le cas $\gamma > 0$ correspond au domaine d’attraction de Fréchet de paramètre $1/\gamma$ appelée aussi loi à queue lourde. On y trouve les lois de Pareto, de Cauchy, de Student, etc. Le cas $\gamma < 0$ correspond au domaine d’attraction de Weibull et le cas $\gamma = 0$ au domaine d’attraction de Gumbel. Cependant la loi limite évoquée ci-dessus ne donne pas d’information sur les autres valeurs extrêmes de l’échantillon du maximum. La Figure 5.4 illustre le comportement de différentes densités de la distribution GEV pour certaines valeurs de γ . Nous présentons dans le Tableau 5.5 une liste de distributions appartenant aux domaines d’attraction de lois GEV. Une deuxième approche a été introduite par Pickands basée sur la distribution limite des valeurs qui dépassent un certain seuil fixé u . On définit un excès par la variable $Y = (X - u)\mathbb{1}_{X > u}$. Lorsque F satisfait le Théorème de Fisher-Tippett-Gnedenko, alors pour u suffisamment grand, la distribution de la variable aléatoire Y est

approximativement

$$H(y) = 1 - \left(1 + \frac{\gamma y}{\sigma + \gamma(u - \mu)}\right)$$

définit sur l'ensemble $\{y : y > 0 \text{ et } 1 + \frac{\gamma y}{\sigma + \gamma(u - \mu)} > 0\}$. La famille de cette loi limite est appelée famille de Pareto généralisée. Quelques estimateurs de l'indice des valeurs extrêmes (IVE) γ existent dans la littérature essentiellement dans le cas i.i.d. La maximisation de la log-vraisemblance de la variable de loi GEV ne conduit pas en général à des solutions analytiques mais pour un ensemble de données la maximisation peut être obtenue à l'aide de méthodes numériques. Des familles d'estimateurs de l'indice de queue de distribution ont été explorées. L'IVE contrôle le comportement des queues de distribution. Son estimation a été largement étudiée dans la littérature.

L'estimateur le plus populaire dans le cas i.i.d et valable pour des distributions appar-

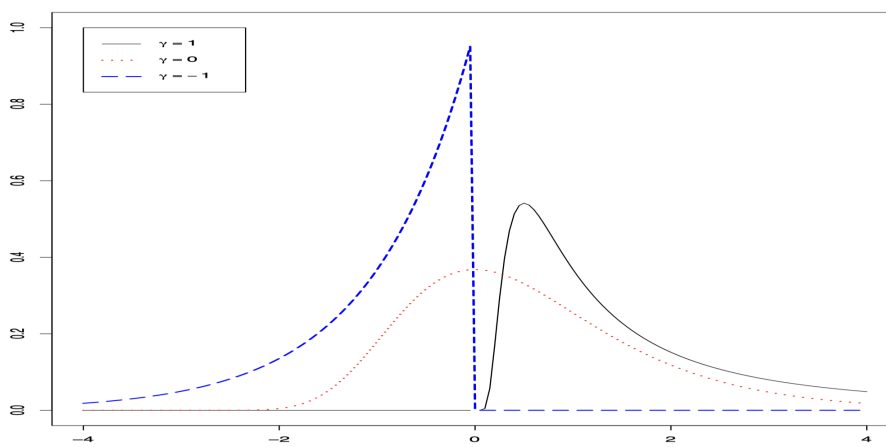


FIGURE 5.4 – Densités de la distribution des valeurs extrêmes pour différentes valeurs de l'indice de queue γ .

tenant au domaine de Fréchet est celui de Hill [2], basé sur les différences entre les logarithmes des plus grandes statistiques d'ordre :

$$\hat{\gamma}_n^H = \frac{1}{k} \sum_{j=1}^k \left(\log X_{n+1-j,n} - \log X_{n-k,n} \right), \quad (5.3)$$

où $X_{k,n}$ est la k -ième statistique d'ordre et $k = k(n)$ vérifie : $k \rightarrow +\infty$, $k/n \rightarrow 0$, $n \rightarrow +\infty$. L'équation (5.3) est une extension de la méthode du maximum de vraisemblance appliquée à la queue de la distribution F . Pour $\gamma > 0$, on vérifie la condition suivante : $\frac{1-F(x)}{1-F(u)} = \left(\frac{x}{u}\right)^{-1/\gamma} \quad \forall x \geq u$. Il est alors raisonnable d'utiliser les statistiques d'ordre qui dépassent le seuil u : $X_{n-k_n+1,n}, \dots, X_{n,n}$ avec une fonction de log-vraisemblance donnée par

$$\begin{aligned} l(X_{n-k_n+1,n}, \dots, X_{n,n}, \gamma) &= -k_n \log(\gamma u) + \log(1 - F(u)) \\ &\quad - \frac{1 + \gamma}{\gamma} \sum_{j=1}^{k_n} \left(\log X_{n-j+1,n} - \log(u) \right). \end{aligned}$$

En remplaçant le seuil u par la statistique d'ordre $X_{n-k_n,n}$ on obtient l'estimateur de Hill. Il existe des estimateurs alternatifs dans le cas indépendant (estimateurs de Pickands, revisité par Dekkers et de Haan,[59],[61]). Peu de travaux ont été développés dans le

Distributions	$1 - F(x)$	γ
Burr(β, τ, λ), $\beta > 0, \tau > 0, \lambda > 0$	$\left(\frac{\beta}{\beta+x^\tau}\right)^\lambda$	$\frac{1}{\lambda\tau}$
Fréchet($1/\alpha$), $\alpha > 0$	$1 - \exp(-x^{-\alpha})$	$\frac{1}{\alpha}$
Loggamma(m, λ) $m > 0, \lambda > 0$	$\frac{\lambda^m}{\Gamma(m)} \int_x^\infty (\log u)^{m-1} u^{-\lambda-1} du$	$\frac{1}{\lambda}$
Loglogistic(β, α), $\beta > 0, \alpha > 1$	$\frac{1}{1+\beta x^\alpha}$	$\frac{1}{\alpha}$
Pareto(α), $\alpha > 0$	$x^{-\alpha}$	$\frac{1}{\alpha}$
Pareto Généralisé (α), $\alpha > 0$	$(1+x/\alpha)^{-\alpha}$	$\frac{1}{\alpha}$
Cauchy	$\int_\infty^x \frac{1}{\pi(1+u^2)} du$	1
Gamma(m, λ), $m > 0, \lambda > 0$	$\frac{\lambda^m}{\Gamma(m)} \int_x^\infty u^{m-1} \exp(-\lambda u) du$	0
Gumbel(μ, β), $\mu \in \mathbb{R}, \beta > 0$	$\exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right)$	0
Logistic	$\frac{2}{1+\exp(x)}$	0
Lognormale(μ, σ), $\mu \in \mathbb{R}, \sigma > 0$	$\frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{u} \exp\left(-\frac{1}{2\sigma^2}(\log u - \mu)^2\right) du$	0
Weibull(λ, τ), $\lambda > 0, \tau > 0$	$\exp(-\lambda x^\tau)$	0
Uniforme(0, 1)	$1 - x$	-1
Beta(α, β)	$1 - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x u^{\alpha-1}(1-u)^{\beta-1} du$	$-1/\beta$
ReverseBurr($\beta, \tau, \lambda, x_F$), $\beta > 0, \tau > 0, \lambda > 0$	$\left(\frac{\beta}{\beta+(x_F-x)^{-\tau}}\right)^\lambda$	$-\frac{1}{\lambda\tau}$

FIGURE 5.5 – Une liste de distributions appartenant aux domaines d'attraction de lois GEV.

cas dépendant (cas de variables aléatoires mélangeantes). Notons que la consistance (en probabilité ou presque) de l'estimateur $\hat{\gamma}_n^H$ ne dépend que du comportement de k , alors que sa normalité asymptotique requiert des conditions supplémentaires sur la fonction de distribution F et donc sur la fonction quantile de la queue de distribution U . Ainsi, une condition dite de second ordre est généralement imposée (Variation régulière du second ordre) :

Condition de second ordre (C_{SO}). Supposons qu'il existe une fonction positive ou négative A vérifiant $\lim_{t \rightarrow \infty} A(t) = 0$ et $\rho < 0$ telle que

$$\lim_{t \rightarrow \infty} \frac{1}{A(t)} \left(\frac{U(tx)}{U(t)} - x^\rho \right) = x^\rho \frac{x^\rho - 1}{\rho}, \quad \forall x > 0,$$

avec $U = \frac{1}{1-F^\leftarrow}$ et F^\leftarrow désignant la pseudo-inverse de F . Le paramètre ρ (appelé paramètre de second ordre) contrôle la vitesse de convergence. La vitesse de convergence de la fonction A vers 0 est importante car permet d'illustrer le terme de biais des estimateurs de l'indice de queue, sous l'hypothèse que $\sqrt{k_n} A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, lorsque $n \rightarrow \infty$ et en supposant les conditions de régularité suivantes sur les coefficients β -mélange :

Condition de régularité (C_R). Il existe $\epsilon > 0$, une fonction bi-variée r et une suite ℓ_n telle que, pour $n \rightarrow \infty$,

$$(a) \quad \frac{\beta(\ell_n)}{\ell_n} n + \ell_n \frac{\log^2 k_n}{\sqrt{k_n}} \longrightarrow 0;$$

$$(b)$$

$$\frac{n}{\ell_n k_n} \text{Cov} \left(\sum_{i=1}^{\ell_n} \mathbb{I}_{\{X_i > F^{\leftarrow}(1-k_n x/n)\}}, \sum_{i=1}^{\ell_n} \mathbb{I}_{\{X_i > F^{\leftarrow}(1-k_n y/n)\}} \right) \longrightarrow r(x, y);$$

avec $0 \leq x, y \leq 1 + \epsilon$.

$$(c) \quad \frac{n}{\ell_n k_n} \mathbb{E} \left[\left(\sum_{i=1}^{\ell_n} \mathbb{I}_{\{F^{\leftarrow}(1-k_n y/n) < X_i \leq F^{\leftarrow}(1-k_n x/n)\}} \right)^4 \right] \leq C(y - x),$$

avec $0 \leq x < y \leq 1 + \epsilon$; $n \in \mathbb{N}$ où C est une constante donnée. On a le résultat suivant de [60] concernant la normalité asymptotique de $\hat{\gamma}_{k_n}^{(H)}$.

Théorème 5.2. *La normalité asymptotique de $\hat{\gamma}_{k_n}^{(H)}$ [60] est donnée comme suit :*

$$\sqrt{k}(\hat{\gamma}_{k_n}^{(H)} - \gamma) \xrightarrow{d} \mathcal{N} \left(\frac{\lambda}{1 - \rho}, \sigma_H^2(\gamma) \right), \quad (5.4)$$

où $\sigma_H^2(\gamma) = \gamma^2 r(1, 1)$ et r la structure de la covariance, avec une expression simple dans le cas i.i.d., où $\sigma_H^2(\gamma) = \gamma^2$.

En pratique, le terme de biais de $\hat{\gamma}_{k_n}^{(H)}$ dépend du fait que ρ est proche de zéro ou non, puisque sous la condition du second ordre (C_{SO}), la fonction $|A|$ varie régulièrement à l'infini avec l'indice ρ . Cela explique toute la littérature consacrée à la correction des biais dans le contexte des données expérimentales par exemple [65], [62] et [63] parmi d'autres. Toutefois, dans le cas des séries de valeurs β -mélangeantes, seuls les articles récents de [59] et [64] traitent de ce problème et proposent un estimateur corrigé du biais pour γ . En outre, ils ont établi les propriétés asymptotiques de γ sous les conditions de régularité et les hypothèses du second ordre et introduit deux classes d'estimateurs de l'indice de queue, qui généralisent dans le cas i.i.d l'estimateur de Hill. En utilisant une méthodologie Jackknife sur deux de leurs estimateurs, ils ont construit des estimateurs asymptotiquement non biaisés pour l'indice de queue γ . Dans [69], nous adaptons cette méthodologie dans le cas dépendant à savoir le cas β mélangeant pour estimer l'indice de valeur extrême ou un quantile extrême. Nous considérons les statistiques introduites dans [61] et définies par :

$$M_{k_n}^{(\alpha)} := \frac{1}{k_n} \sum_{i=1}^{k_n} (\log X_{n-i+1, n} - \log X_{n-k_n, n})^\alpha, \quad \alpha > 0.$$

Ceci peut être réécrit comme une fonction de $(Q_n(t) := X_{n-[k_n t], n})_{t \in [0, 1]}$, et le processus quantile de queue comme suit

$$M_{k_n}^{(\alpha)} = \int_0^1 \left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha dt.$$

Proposition 5.1. *Supposons que (X_1, X_2, \dots) est une série temporelle stationnaire β mélangée avec une fonction de distribution marginale commune continue F et supposons que les conditions (C_{SO}) et (C_R) sont vérifiées. Soit k_n une suite telle que $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ et $k_n^{1/2}A(n/k_n) = O(1)$, avec $n \rightarrow \infty$. Alors, pour $\alpha, \delta > 0$, il existe une fonction $\tilde{A} \sim A$, et un processus gaussien centré $(W(t))_{t \text{ dans } [0,1]}$ avec une fonction de covariance $r(\cdot, \cdot)$ définie dans des conditions de régularité (C_R) , tel que, pour $n \rightarrow \infty$:*

$$\sup_{t \in (0,1]} t^{1/2+\delta} \left| \sqrt{k} \left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha - \gamma^\alpha (-\log t)^\alpha - \alpha \gamma^\alpha (-\log t)^{\alpha-1} (t^{-1}W(t) - W(1)) - \sqrt{k} \tilde{A}(n/k) \alpha \gamma^{\alpha-1} (-\log t)^{\alpha-1} \frac{t^{-\rho} - 1}{\rho} \right| \rightarrow 0 \text{ p.s.}$$

Le théorème suivant donne les développements asymptotiques des estimateurs $\tilde{\gamma}_{k_n}^{(\alpha)}$ et $\hat{\gamma}_{k_n}^{(\alpha)}$ en termes de processus gaussien $P^{(\alpha)}$ où

$$P^{(\alpha)} = \int_0^1 (-\log t)^{\alpha-1} (t^{-1}W(t) - W(1)) dt$$

est une variable aléatoire normalement distribuée de moyenne nulle et de covariance $Cov(P^{(\alpha)}, P^{(\alpha)}) = c$, définie comme suit,

$$c_{\alpha,\beta} = \int_0^1 \int_0^1 (-\log s)^{\alpha-1} (-\log t)^{\beta-1} \times \left(\frac{r(s,t)}{st} - \frac{r(s,1)}{s} - \frac{r(1,t)}{t} + r(1,1) \right) ds dt,$$

avec la structure de covariance r définie comme dans la condition de régularité (C_R) .

Théorème 5.3. *Supposons que les conditions de la proposition 5.1 soient remplies. On a alors, comme $n \rightarrow \infty$:*

$$\tilde{\gamma}_{k_n}^{(\alpha)} \stackrel{d}{=} \gamma + \frac{\gamma P^{(\alpha)}}{k_n^{1/2} \Gamma(\alpha+1)} + \tilde{A}(n/k_n) \frac{1 - (1-\rho)^\alpha}{\alpha \rho (1-\rho)^\alpha} + o_{\mathbb{P}}(k_n^{-1/2}) \text{ for } \alpha > 0 \quad (5.5)$$

et pour $\alpha \geq 1$

$$\hat{\gamma}_{k_n}^{(\alpha)} \stackrel{d}{=} \gamma + \frac{\gamma \alpha P^{(\alpha)}}{k_n^{1/2} \Gamma(\alpha+1)} - \gamma(\alpha-1) \frac{P^{(1)}}{k_n^{1/2}} + \tilde{A}(n/k_n) \left\{ \frac{1 - (1-\rho)^\alpha}{\rho(1-\rho)^\alpha} - \frac{\alpha-1}{1-\rho} \right\} + o_{\mathbb{P}}(k_n^{-1/2}). \quad (5.6)$$

Maintenant, nous présentons notre classe d'estimateurs à biais réduit en utilisant l'approche Jackknife. Les estimateurs de l'indice de queue qui en résultent sont les suivants :

$$\hat{\gamma}_{k_n}^{(H)} = \tilde{\gamma}_{k_n}^{(1)} = \hat{\gamma}_{k_n}^{(1)} = M_{k_n}^{(1)}, \quad \tilde{\gamma}_{k_n}^{(2)} = \sqrt{\frac{M_{k_n}^{(2)}}{2}}. \quad (5.7)$$

À partir de n'importe quelle paire de ces statistiques, nous construisons les estimateurs adaptatifs suivants de l'indice de queue à biais réduit dans le sens de la méthodologie Jackknife généralisée :

$$\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub1)} := \frac{1}{\hat{\rho}} \left(\sqrt{2M_{k_n}^{(2)}} - (2 - \hat{\rho}) \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \right),$$

$$\widehat{\gamma}_{k_n, \widehat{\rho}}^{(Ub2)} := \frac{1}{\widehat{\rho}} \left((2 - \widehat{\rho}) M_{k_n}^{(1)} - (1 - \widehat{\rho}) \sqrt{2M_{k_n}^{(2)}} \right)$$

et

$$\widehat{\gamma}_{k_n, \widehat{\rho}}^{(dH)} := \frac{1}{\widehat{\rho}} \left(M_{k_n}^{(1)} - (1 - \widehat{\rho}) \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \right),$$

où $\widehat{\rho}$ est soit une valeur négative canonique $\widehat{\rho} := \rho = \rho_0$ soit un estimateur consistant $\widehat{\rho} := \widehat{\rho}_{k_\rho}$ de ρ , avec $k_\rho := k_{\rho, n}$ une suite intermédiaire d'entiers supérieurs à k_n , satisfaisant $k_\rho \rightarrow \infty$ et $k_\rho/n \rightarrow 0$, avec $n \rightarrow \infty$. Notons que l'estimateur $\widehat{\gamma}_{k_n, \widehat{\rho}_{k_\rho}}^{(dH)}$ peut être écrit comme suit :

$$\widehat{\gamma}_{k_n, \widehat{\rho}_{k_\rho}}^{(dH)} = M_{k_n}^{(1)} - \frac{M_{k_n}^{(2)} - 2 \left(M_{k_n}^{(1)} \right)^2}{2M_{k_n}^{(1)} \widehat{\rho}_{k_\rho} \left(1 - \widehat{\rho}_{k_\rho} \right)^{-1}}.$$

Théorème 5.4. *Soit (X_1, X_2, \dots) une série temporelle stationnaire β -mélangeante avec une fonction de distribution marginale commune continue F et supposons que (C_{SO}) et (C_R) sont véridiques. Soit k_n une suite intermédiaire telle que $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ et $\sqrt{k_n} A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, avec $n \rightarrow \infty$. Si $\widehat{\rho}$ est soit une valeur négative canonique $\widehat{\rho} := \rho = \rho_0$ ou un estimateur $\widehat{\rho} := \widehat{\rho}_{k_\rho}$ de ρ , consistant avec $k_\rho := k_{\rho, n}$ une autre suite intermédiaire d'entiers supérieurs à k_n et satisfaisant $k_\rho \rightarrow \infty$ et $k_\rho/n \rightarrow 0$, avec $n \rightarrow \infty$. Alors on a :*

$$\sqrt{k_n} \left(\widehat{\gamma}_{k_n, \widehat{\rho}}^{(Ub1)} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\gamma, \rho) \right),$$

$$\sqrt{k_n} \left(\widehat{\gamma}_{k_n, \widehat{\rho}}^{(Ub2)} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\gamma, \rho) \right)$$

et

$$\sqrt{k_n} \left(\widehat{\gamma}_{k_n, \widehat{\rho}}^{(dH)} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\gamma, \rho) \right),$$

avec

$$\sigma^2(\gamma, \rho) = \frac{\gamma^2}{\rho} \left((2 - \rho)^2 c_{1,1} + (1 - \rho)^2 c_{2,2} + 2(2 - \rho)(\rho - 1)c_{1,2} \right).$$

Nous illustrons les performances de notre procédure de réduction des biais à travers une étude de cas réel portant sur les données financières de l'indice *S&P500*. Pour cela, nous appliquons l'indice extrême et les quantiles extrêmes pour évaluer l'indice *S&P500* : $(I_t, t = 1, \dots, n)$. Les données de la figure 5.6 montrent les retours journaliers négatifs quotidiens $X_t = \log(I_t/I_{t-1})$, (*Loss*) pour $n = 1000$ valeurs de l'indice *S&P500* du *avril 23th*, 2018 à *April 8th*, 2022. Dans une perspective de gestion des risques, la valeur à risque (*VaR*) est une quantité courante inscrite dans le cadre réglementaire international appelé *Accords de Bâle*. L'Accord de Bâle exige que les plus grandes banques internationales détiennent des fonds propres réglementaires pour leur portefeuille de négociation sur la base d'une *VaR* de 99% sur une période de détention de 1 ou 10 jours. Le calcul du capital-risque basé sur la *VaR* a été étudié au cours des deux dernières décennies, voir [69] pour plus de détails.

Par ailleurs dans [85], nous proposons des estimateurs à biais réduit pour des distributions de pertes à queues lourdes X_i (coûts des sinistres) d'un modèle d'assurance appelé "probabilité de ruine à temps infini". Le risque de ruine est un concept clé dans le domaine de l'assurance. Il désigne la probabilité que la réserve financière d'une compagnie d'assurance, calculée comme la différence entre les primes collectées et les montants payés pour les réclamations, devienne négative à un moment donné. Pour établir la probabilité

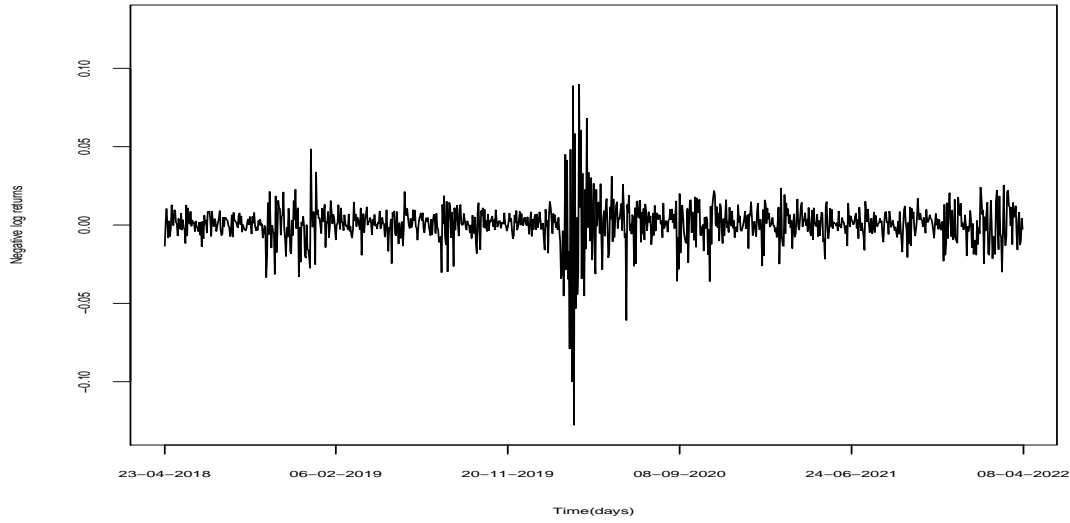


FIGURE 5.6 – Données d'index *S&P500* : résultats de journal négatifs quotidiens du 23 avril 2018 au 8 April 2022.

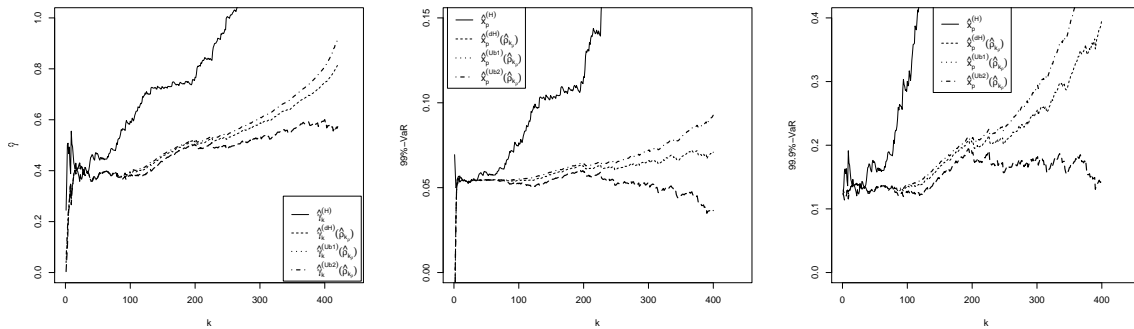


FIGURE 5.7 – Données de l'indice *S&P500* : valeurs estimées de γ (panel de gauche), 99%-VaR (panel du milieu) et 99,9%-VaR (panel de droite). panel) en fonction de k_n .

de ruine en temps infini, on suppose d'abord que les variables aléatoires X_i sont indépendantes de M_t (Le nombre de sinistres sur la période $]0, t]$) et que l'assureur collecte une prime à un taux constant c par unité de temps, en satisfaisant la condition de profit net. Le processus de risque classique $(R_t)_{t>0}$ est donné par :

$$R_t := u + ct - \sum_{i=1}^{M_t} X_i, \quad t > 0; \quad u = R_0 > 0 \text{ (la réserve initiale).}$$

Le processus correspondant de surplus de réclamation est défini par :

$$S_t := u - R_t = ct - \sum_{i=1}^{M_t} X_i.$$

Tout d'abord, nous nous intéressons à la probabilité que S_t dépasse une réserve initiale u à un moment t avant ou à un horizon T . Explicitement, cette probabilité peut s'écrire comme suit : $\Psi_T(u) := \mathbb{P} \left\{ \sup_{0 \leq t \leq T} S_t > u \right\}$ et la probabilité de ruine à temps

infini $\phi(u) = \lim_{T \rightarrow \infty} \Psi_T(u)$. En sciences actuarielles, les coûts élevés liés à de grosses pertes nécessitent une modélisation des événements ayant une faible probabilité de se produire (événements extrêmes). L'analyse de ces événements extrêmes peut être réalisée en utilisant la méthodologie des valeurs extrêmes. Ces sujets d'étude ont attiré beaucoup d'attention dans la littérature. Cependant, les estimations disponibles souffrent fortement de sous-estimation ou ont un problème de robustesse, en particulier lorsque les pertes sont perturbées par de grandes variations (valeurs aberrantes). Ainsi dans l'article [85], nous introduisons un estimateur robuste de la probabilité de ruine en temps infini pour de telles distributions. Notre méthodologie est basée sur les estimateurs dits de t-Hill (t-score ou score moment estimation) pour l'indice de toute queue de distribution. Nous établissons leur normalité asymptotique et, par le biais d'une étude de simulation, nous illustrons leur comportement en termes de biais absolu et d'erreur quadratique moyenne. Les résultats de la simulation montrent clairement que nos estimateurs sont performants et qu'ils sont assez robustes aux valeurs aberrantes. D'autre part dans [70], on s'intéresse à la queue de distribution des revenus dans de nombreux pays en étudiant le ratio de la part des quintiles (QSR), une mesure d'inégalité des revenus récemment introduite dans la littérature scientifique, faisant également partie des indicateurs européens de Laeken et qui couvre quatre dimensions importantes de l'inclusion sociale (la santé, l'éducation, l'emploi et la pauvreté financière). Une estimation non paramétrique a été développée sur l'indice QSR pour les distributions de revenus du capital à queue lourde. Cependant, cette méthode d'estimation ne donne pas de performances statistiques satisfaisantes. Nous avons développé des estimateurs de type semi-paramétrique de l'indice QSR pour les distributions de revenus à queue lourde. Notre méthodologie est basée sur la théorie des valeurs extrêmes. Nous établissons leur distribution asymptotique et, par le biais d'une étude de simulation, nous illustrons leur comportement en termes de biais absolu et d'erreur quadratique médiane. Les résultats de simulation montrent clairement que nos estimateurs fonctionnent bien. D'autres travaux ont été menés [89] et des perspectives de recherche en lien avec l'estimation des paramètres des lois α -stables qui sont asymptotiquement du domaine d'attraction de Pareto [97].

Modélisation spatiale appliquée et apprentissage automatique

Cette partie de ma présentation s'intéresse aux modèles spatiaux et spatio-temporels appliqués aux données environnementales et de santé publique à Mayotte. Le territoire de Mayotte et la situation environnementale et sanitaire observée en relation avec les interactions avec le milieu constituent un terrain favorable particulièrement exposé aux risques sanitaires et sismiques. Par ailleurs, l'île de Mayotte est entourée par l'un des plus grand lagon au monde et par une barrière récifale s'étendant jusqu'à 15 km de large et 70 km de fond formant ainsi l'une des plus grandes et variées formations récifales de l'océan Indien.

6.1 Modélisation spatiale de vecteurs d'arboviroses

Dans le cadre de la modélisation spatiale en santé-environnement, un travail a été mené en collaboration avec la service de lutte anti-vectorielle de l'Agence Régionale de Santé de Mayotte. Il s'agit d'un projet de modélisation pour la surveillance des arboviroses vecteurs de nombreux agents pathogènes, notamment les populations de moustiques vecteurs de la dengue, chikungunya et zika. L'humanité subit des nuisances importantes et des maladies transmises par des moustiques. Les conditions climatiques et environnementales sont connues pour être des déterminants de leur abondance et leur capacité à transmettre ces agents pathogènes. Des données temporelles et spatialisées concernant le cycle de vie du vecteur *Aedes albopictus* aux stades oeufs, larves et adultes sont collectées par le service de lutte anti-vectorielles. Le moyen le plus efficace pour surveiller ces arboviroses est de comprendre et lutter contre leur développement.

Analyse de données réelles

Les données de cette étude concerne les comptages des oeufs et adultes de 2018 à 2021 collectées sur des sites géographiques par jour ou par semaine sur l'île de Mayotte par le service de lutte anti-vectorielle de l'ARS.

Pour chaque stade (oeuf, adulte) une analyse et nettoyage des données ont été faites ainsi que des fusions avec les données environnementales (température, précipitations et

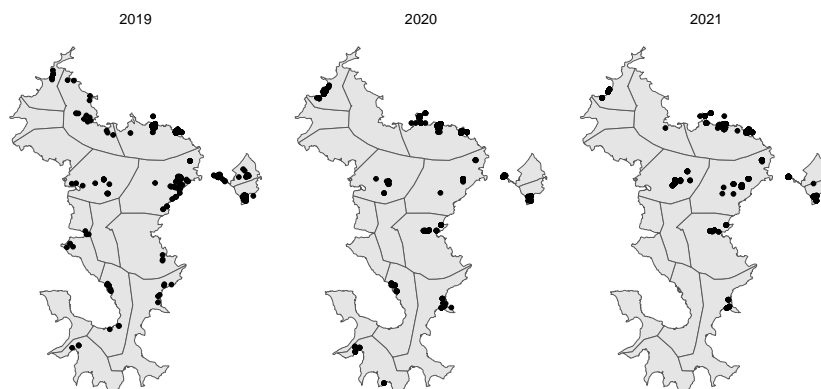


FIGURE 6.1 – Répartition spatiale globale des sites de collecte des oeufs de moustiques de 2019 à 2021.

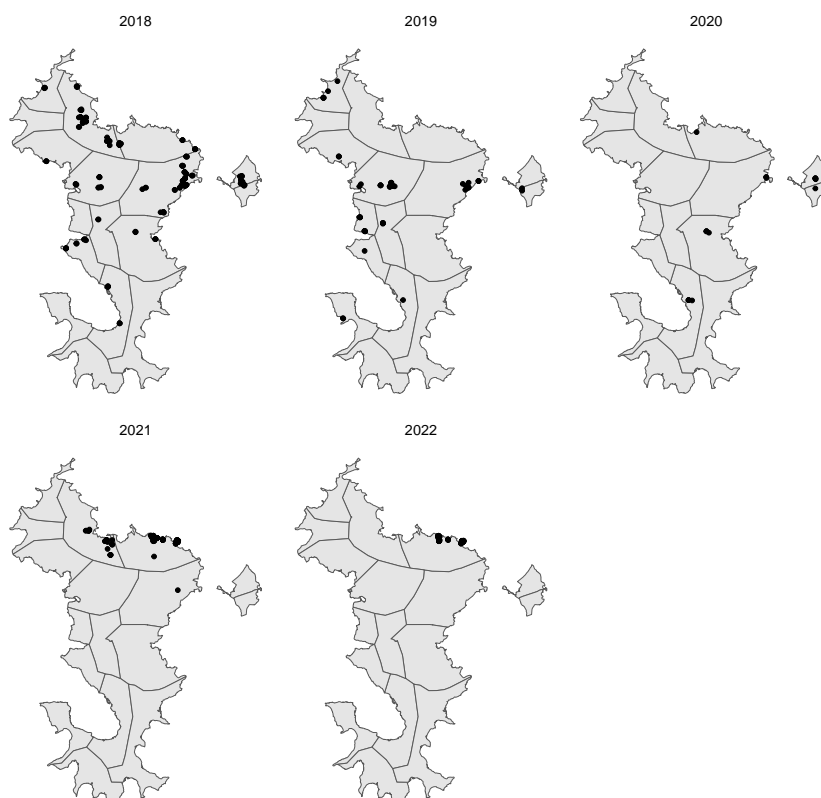


FIGURE 6.2 – Répartition spatiale globale des sites de collecte de moustiques adultes de 2019 à début 2022.

les déchets artificiels comme logements insalubres, des dépôts d'eaux usées, des déchets de pneus, des débris) par secteur géographique et en fonction du jour et de la semaine.

De telles données covariables sont des facteurs bien connus dans la prolifération des populations de moustiques. Une grille 100×100 de la carte de Mayotte a été réalisée pour imputer ces co-variables et les données de capture dans chaque commune ou vil-

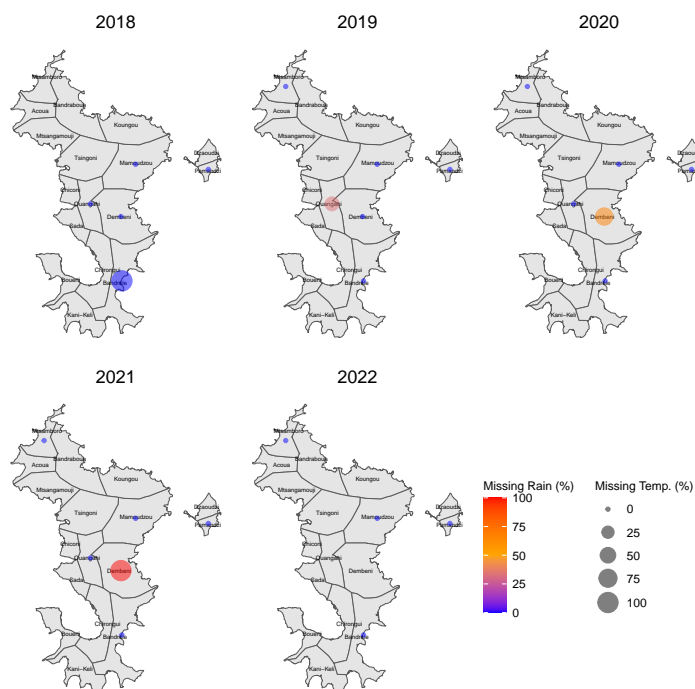


FIGURE 6.3 – Carte des stations météorologiques de Mayotte et données manquantes

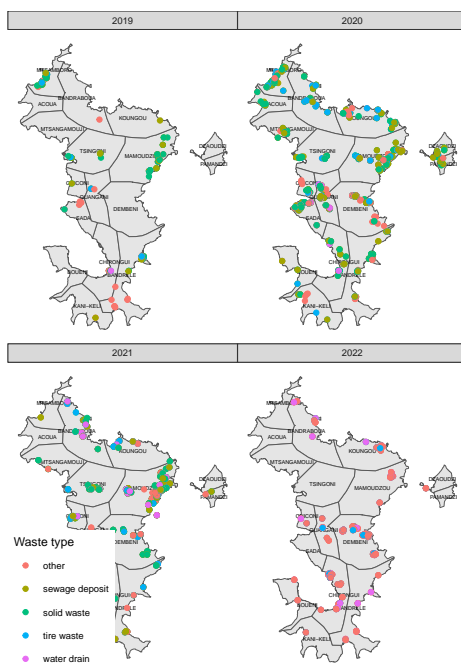


FIGURE 6.4 – Distribution annuelle des déchets artificiels à Mayotte

lage à partir des observations voisines en utilisant la méthode des k plus proches voisins (imputation spatiale) et les méthodes de quantiles et de regression. La qualité de ces méthodes d'imputation est discutée.

Nous avons proposé une modélisation spatio-temporelle pour la surveillance de l'abon-

dance du nombre d'oeufs et de moustiques adultes, en tenant compte des données environnementales (y compris l'introduction de retard) dans le contexte de l'épidémiologie des moustiques et les données de reproduction artificielle. Après une analyse préliminaire de l'ensemble des données, nous avons jugé nécessaire la mise en place de méthodes de régression type Lasso pour éliminer la multicollinéarité potentielle entre nos multiples variables explicatives. Des études préliminaires de modélisation statistique ont confirmé le choix d'un Modèle Additif Généralisé (GAM) dans un cadre spatio-temporel qui peut expliquer l'abondance des moustiques adultes et oeufs. Il met en évidence les facteurs pertinents pour prendre en compte les interactions non linéaires et les effets de la surdispersion. Nous évaluons le résultats des prédictions spatio-temporelles, par l'introduction d'un ratio d'abondance spatio-temporel $R(\mathbf{s}, t)$ au point spatial \mathbf{s} sur une grille spatiale $\mathbf{s} \in \{\mathbf{s}_1, \dots, \mathbf{s}_m\}$ de taille m et au temps t mesuré sur une échelle hebdomadaire $t \in \{t_1, \dots, T\}$, défini comme suit [88] :

$$R(\mathbf{s}, t) = \frac{\hat{Y}(\mathbf{s}, t) - \min_{\mathbf{s}, t}(\hat{Y}(\mathbf{s}, t))}{\max_{\mathbf{s}, t}(\hat{Y}(\mathbf{s}, t)) - \min_{\mathbf{s}, t}(\hat{Y}(\mathbf{s}, t))} \quad (6.1)$$

où $\hat{Y}(\mathbf{s}, t)$ est l'abondance en temps et espace du moustique (au stade oeuf ou adulte). La quantité $R(\mathbf{s}, t)$ fournit des probabilités d'abondance des moustiques et permet de mettre en évidence les zones spatio-temporelles potentiellement à risque d'abondance où les services de santé pourraient appliquer des mesures de surveillance et de contrôle des vecteurs de moustiques. Une variété de modèles d'apprentissage des données et de prédiction y sont déployés : la régression Lasso pour la stratégie de sélection des variables (étant donné que nous sommes en présence d'un grand nombre de variables explicatives), un modèle additif généralisé (Gam) et un modèle à noyau spatio-temporel. Les deux premiers modèles sont des classiques de la littérature scientifique et statistique. Nous décrivons le troisième modèle comme suit. Un prédicteur spatio-temporel à noyaux d'une population de moustiques \hat{Y} au point spatial non observé \mathbf{s}_0 et au temps t_0 , est donné comme suit :

$$\hat{Y}(\mathbf{s}_0; t_0) = \sum_{j=1}^T \sum_{i=1}^m K_{ij}(\mathbf{s}_0; t_0) Y(\mathbf{s}_i, t_j) \text{ avec } K_{ij}(\mathbf{s}_0; t_0) = \frac{\tilde{K}_{ij}(\mathbf{s}_0; t_0)}{\sum_{k=1}^T \sum_{l=1}^m \tilde{K}_{lk}(\mathbf{s}_0; t_0)},$$

où

$$\tilde{K}_{ij}(\mathbf{s}_0, t_0) = \frac{1}{d(\mathbf{s}_i; t_j), (\mathbf{s}_0; t_0)^\theta} \quad (6.2)$$

avec d une distance entre les points spatio-temporels $(\mathbf{s}_i; t_j)$ et (\mathbf{s}_0, t_0) . Le noyau de transition K dépend d'un paramètre de lissage θ qui a une fonction de redistribution des poids du processus observé. Ce modèle est basé sur une méthode de pondération par l'inverse de la distance (IDW) donnant plus de poids aux positions les plus proches. Un autre exemple de noyaux est le noyau Gaussien :

$$\tilde{K}_{ij}(\mathbf{s}_0, t_0) = \exp\left(-\frac{1}{\theta} d((\mathbf{s}_i; t_j), (\mathbf{s}_0; t_0))^2\right)$$

où θ est par exemple proportionnel à la variance. D'autres types de noyaux existent dans la littérature mais nous avons choisis pour des raisons d'ordre pratique de nous concentrer sur les deux noyaux précédents. L'objectif visé est d'améliorer la structure de ce modèle spatio-temporel ci-dessus en incluant des variables importantes sélectionnées par des méthodes de régression statistiques adéquats (Lasso, Gam, etc) dans le calcul de

la distance d . Dans ce sens, supposons qu'aux emplacements des points spatiaux $\{\mathbf{s}_i : i = 1, \dots, m\}$ nous relevons k valeurs de caractéristiques associées $c_{ij}^{(1:k)} = (c_{ij}^{(1)}, \dots, c_{ij}^{(k)})$ au temps t_j correspondant à des mesures de variables explicatives importantes pouvant expliquer l'abondance de la population de moustiques. Plus précisément, on peut écrire par exemple

$$\tilde{K}_1((\mathbf{s}_i; t_j), (\mathbf{s}_0; t_0)) = \frac{1}{d((\mathbf{s}_i, c_{ij}^{(1:k)}; t_j), (\mathbf{s}_0, c_0^{(1:k)}, t_0))^\theta}. \quad (6.3)$$

Ce modèle modifié peut être considéré comme une moyenne pondérée des points de données, donnant aux emplacements et valeurs de variables explicatives les plus proches des poids importants tandis que les observations distantes dans l'espace ou dans les caractéristiques auront des poids relativement faibles. Les propriétés théoriques de ce modèle modifié fait l'objet d'un travail en cours. Cependant notons dans le cas du noyau dans (6.2) que les données peuvent conduire à de très petites valeurs proches de zéro de la distance, ce qui pose des problèmes d'explosion dans l'équation ci-dessus. De plus il faudrait également un choix pertinent de cette distance de manière à prendre en compte les effets des caractéristiques dans le temps et dans l'espace. Nous considérons dans ce travail la distance min-max et une transformation des données (incluant les co-variables) sous la même échelle et permettant de traiter le cas des valeurs nulles. Les données disponibles, collectées par l'ARS de Mayotte, présentent beaucoup de données manquantes à la fois en temps et en espace. La Figure 6.6 montre un exemple de prédiction de risque d'abondance spatio-temporel (par semaine) des oeufs de moustiques à l'île de Mayotte.

En terme de travaux de recherche théorique et appliqué incluant l'estimation de paramètres, un travail est en cours en collaboration avec Guilherme Hilario Monteiro, Bedreddine Ainseba (Bordeaux) et Stefanella Boatto (Brésil). Il s'agit d'un modèle spatio-temporel issu d'un modèle différentiel structuré en âge (modèle mécaniste type EDP) pour la simulation d'une population de moustiques. Nous envisageons sur le plan pratique le déploiement d'une application R Shiny dénommée May'Aedes qui permettra de mettre en place un outil flexible et efficace pour étudier le risque d'abondance spatio-temporel des populations de moustiques prédites à l'île de Mayotte.

6.2 Modélisation spatiale de données environnementales

Toujours dans un cadre de modélisation spatiale, j'ai mené un travail de modélisation des données sismiques issues de la crise sismo-volcanique de 2018 à 2021 dans l'est de Mayotte, voir Figure 6.5. Nous avons mis en évidence dans [84] des modèles basés sur les processus spatiaux et spatio-temporels et les séries temporelles. L'objectif étant de comprendre la dynamique de l'évolution spatiale de l'activité sismique à Mayotte. Nous mettons en évidence certains modèles permettant d'illustrer la dynamique spatiale comme les méthodes de clustering non supervisées, l'estimation des densités spatiales des processus spatiaux ponctuels et la modélisation spatio-temporelle avec l'approche décrite ci-dessus par un prédicteur spatio-temporel à noyau. Leurs performances ont été efficaces pour rechercher et comprendre la variation spatiale et temporelle de ces données spatiales principalement générées par le volcan sous-marin détecté autour de Mayotte appelé **Fani Maoré**.

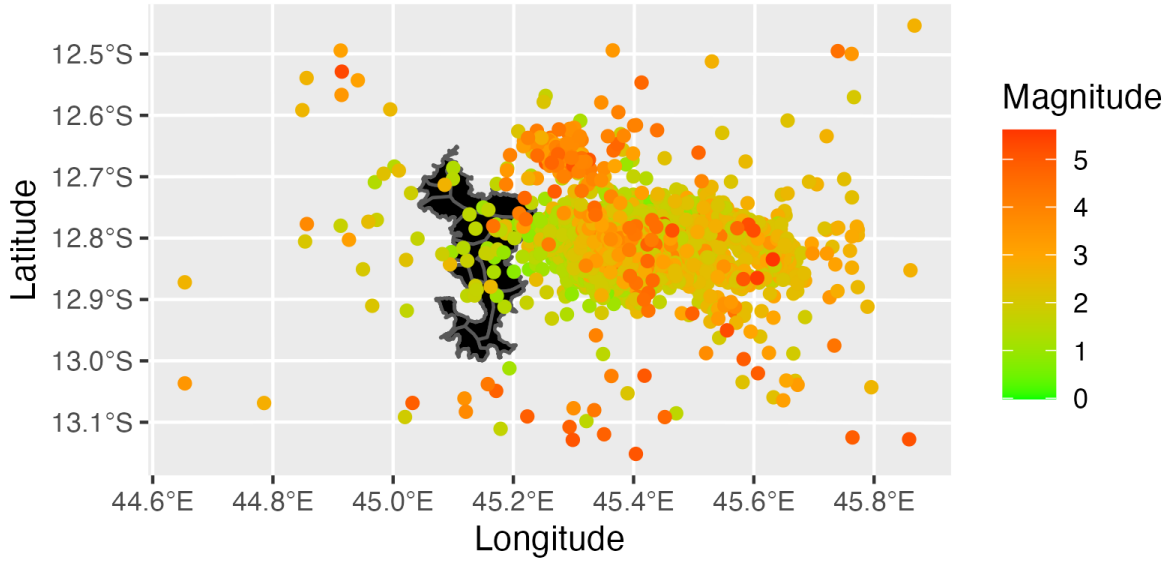


FIGURE 6.5 – La variation spatiale globale observée des données sismiques de 2018 à 2021. (Manou-Abi S.M et al. [84])

Décrivons en quelques lignes une approche spatiale basée sur l'utilisation d'un processus ponctuel marqué et inhomogène en espace [58]. Soit

$$Y = \{(x_1, m_1), \dots, (x_n, m_n)\}, x_i \in W, \quad m_i \in M$$

un processus ponctuel et inhomogène marqué de motif de points marqués où $W \subset \mathbb{R}^2$ désigne la zone d'étude et M l'ensemble des valeurs marquées (par exemples les valeurs des magnitudes). La fonction K de Ripley, pour un modèle spatial de points $\{x_1, \dots, x_n\}$ observé dans une fenêtre plane $W \subset \mathbb{R}^2$ est un outil exploratoire qui peut être utilisé pour évaluer la dépendance entre des emplacements à plusieurs distances. Il est défini par $K(s) = \lambda^{-1} \mathbb{E}(\text{nombre d'événements futurs à une distance } s \text{ d'un événement arbitraire})$ où λ est la fonction d'intensité du processus ponctuel spatial. Pour déterminer si les valeurs d'une fonction K sont relativement grandes ou petites, on compare la fonction K pour le schéma spatial observé avec la fonction K d'un processus de Poisson homogène (CSR) qui est donné par $K(s) = \pi s^2$. Des tests statistiques (par exemple de ξ^2 sur la base d'un comptage par quadrat) existent également pour évaluer la nature de la distribution spatiale observée (Régulière, groupée ou aléatoire). Dans le cas d'une répartition inhomogène, désignons par $\lambda(x, m)$ défini sur $\mathbb{R}^2 \times M$ l'intensité jointe pour les points spatiaux $x \in \mathbb{R}^2$ de marque $m \in M$. La densité de probabilité f du processus ponctuel marqué de Poisson inhomogène Y d'intensité jointe $\lambda(x, m)$ est donnée par :

$$f(y) = \exp\left(\sum_{m \in M} \int_W (1 - \lambda(x_i, m)) dx\right) \prod_{i=1}^{n_m(y)} \lambda(x_i, m)$$

où $n_m(y)$ est le nombre de points $y \in W$ ayant une valeur de marque m . La fonction de densité définit la probabilité d'observer un événement à un endroit $y \in W$ et s'intègre

à l'échelle de la zone d'étude. Les méthodes d'estimation par maximum de vraisemblance s'appliquent. L'estimateur à noyau de Nadaraya-Watson des tendances spatiales des marques lissées à l'emplacement $u \in \mathbb{R}^2$ est donné par

$$\hat{m}(u) = \frac{\sum_i m_i K(u - x_i)}{\sum_i K(u - x_i)}$$

où K désigne un noyau et m_i valeur marquée au point spatial x_i . Les densités spatiales prédites par ce modèle spatial sont données par les Figures 6.7 à 6.8.

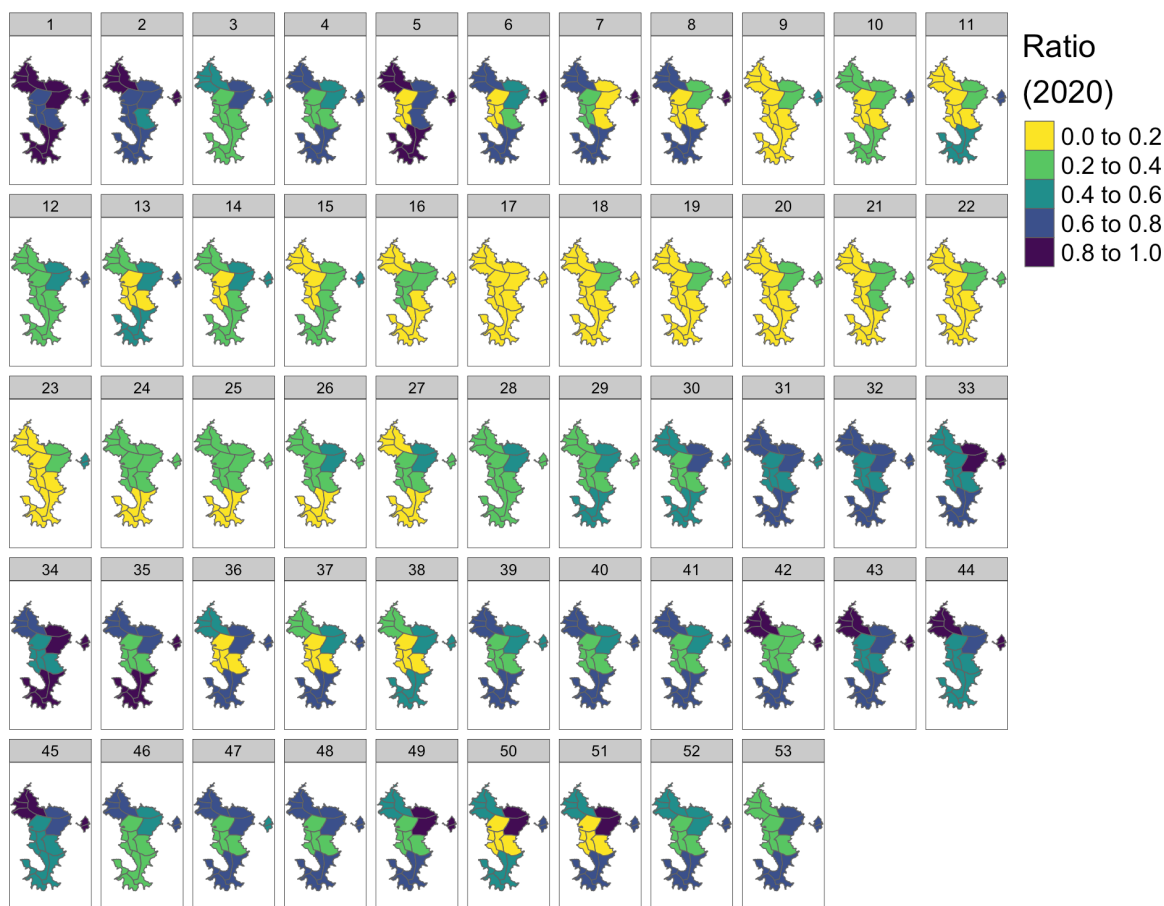


FIGURE 6.6 – Un exemple de prédiction d'abondance spatio-temporel des oeufs de moustiques en 2020 (Manou-Abi S.M et al. [88])

L'article [84] discute également des performances des modèles de Holt-Winters pour illustrer les tendances et saisonnalités ainsi que de nouvelles perspectives de recherche à venir en terme de modélisation mathématique.

Dans un cadre de modélisation spatiale en biologie marine au sein de l'axe [E], on s'intéresse à la production de cartes géomorphologiques marines du banc du Geysier situé entre Mayotte (125 km) et Madagascar (200 km), constituant une ressource patrimoniale dans l'océan indien. Ces dernières sont utiles pour comprendre la structure de ce fond marin et la gestion des ressources ou de la planification de la conservation. Bien que les techniques de construction de ces cartes soient de plus en plus sophistiquées, les techniques manuelles sont encore largement utilisées. Des approches automatisées sont nécessaires pour

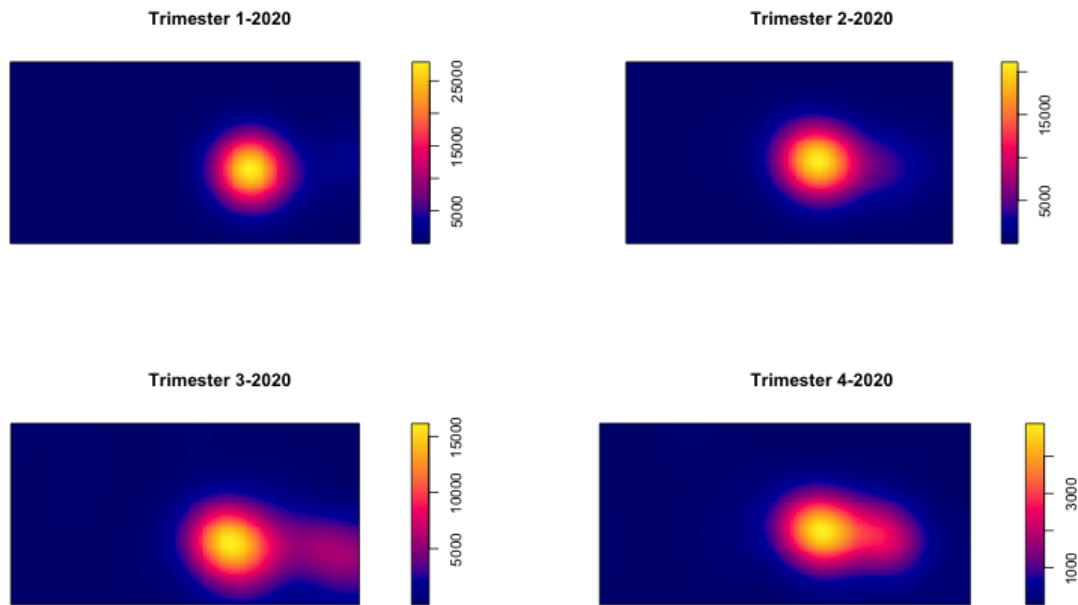


FIGURE 6.7 – Ajustement de la densité spatiale par un processus de Poisson inhomogène marqué au cours de l’année 2020 (Manou-Abi S.M et al. [84])

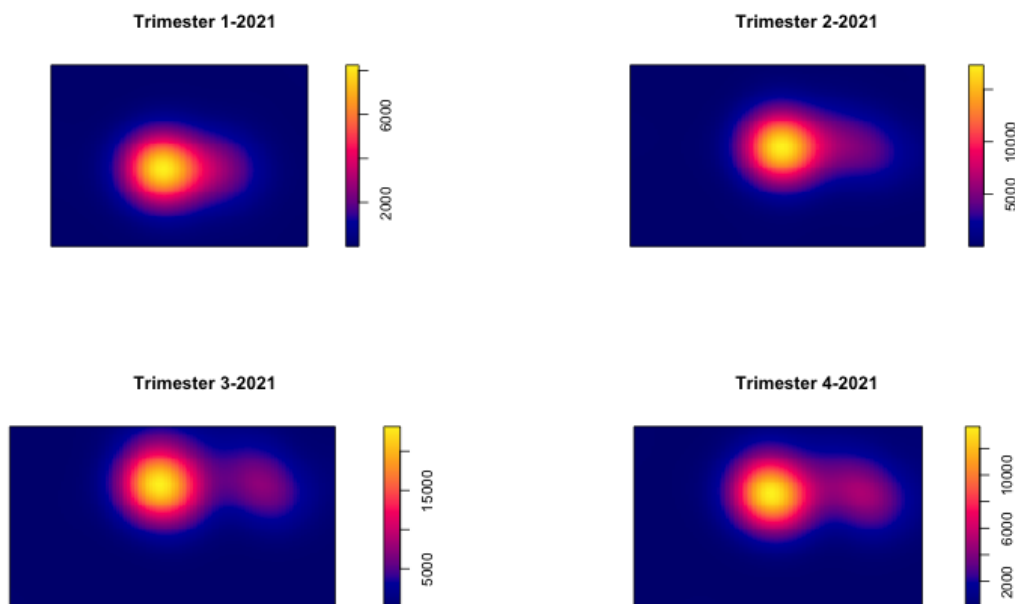


FIGURE 6.8 – Ajustement de la densité spatiale par un processus de Poisson inhomogène marqué au cours de l’année 2021 (Manou-Abi S.M et al. [84])

obtenir des cartes reproductibles en un temps raisonnable. Ainsi dans [92], nous avons développé des approches d’apprentissage statistique de type Machine Learning (algorithmes d’échantillonnage, les forêts aléatoire, interpolation spatiale, méthodes supervisées et non

supervisées, etc) pour construire automatiquement des cartes géomorphologiques. Les résultats ont montré que le cadre proposé permettait de construire efficacement des cartes géomorphologiques pertinentes des fonds marins. Le meilleur modèle (basé sur l'échantillonnage et les forêts aléatoires) a permis d'obtenir une carte correspondant à 90% à la carte de référence dite carte experte.

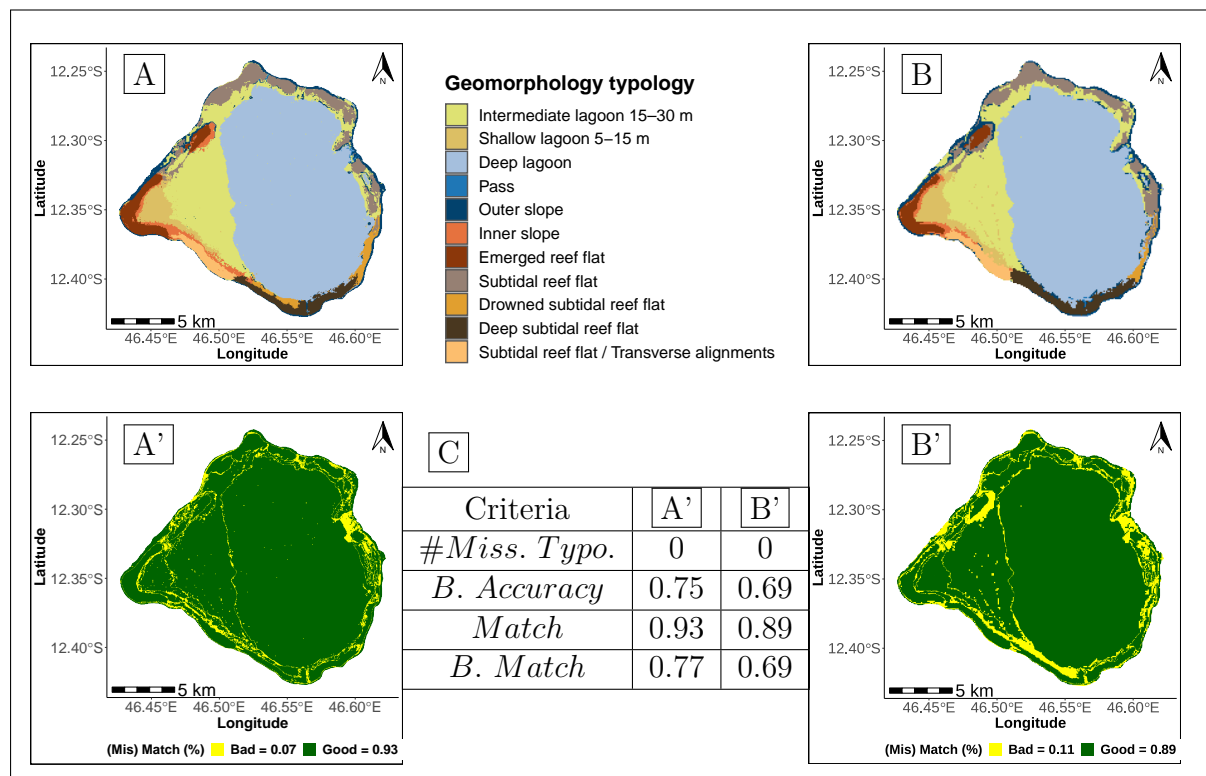


FIGURE 6.9 – A) Carte prédite avec une résolution de 50m sur mille 1000 emplacements échantillonnés. B) Carte prédite avec une résolution de 100 m sur 200 emplacements échantillonnés. Cartes de correspondance/inadéquation correspondantes (A' et B') par rapport à celle des experts. C) Critères de performance mesurés (Faye P. et al. [92])

Projets de recherche en cours

Mon projet de recherche s'intitule **Modélisation, estimation et application aux données réelles : lois et processus stables, modèles stochastiques, spatiaux et apprentissage machine**. Ce projet est structuré en trois grands axes principaux.

Le premier axe, intitulé *Modèles stochastiques : Estimation, Approximation et Apprentissage automatique (Deep Learning)*, présenté en Section 7.1, est en parfaite cohésion avec mes travaux sur les modèles d'équations différentielles stochastiques dirigées par des processus de Lévy, incluant les processus α -stables, ainsi que sur les modèles basés sur des extensions de chaînes de Markov. Les aspects d'estimation non paramétrique et d'approximation numérique ainsi que les approches de Machine Learning seront explorés, de même que les propriétés d'ergodicité et de mélange, en utilisant des approches de couplage en distances de Wasserstein et variation totale. Ce projet sera déployé à moyen et long terme, incluant le développement de packages R ou Python pour l'estimation des coefficients de dérive et de diffusion.

Le second axe, intitulé *Estimation des paramètres des lois stables à l'aide de la théorie des valeurs extrêmes*, présenté en Section 7.2, s'inscrit dans la continuité de mes travaux sur l'estimation statistique associée aux lois à queues lourdes et à la théorie des valeurs extrêmes. Ce projet, actuellement en cours de réalisation, se déroulera à court terme.

Le dernier axe, intitulé *Estimation statistique, modèles spatiaux et apprentissage automatique*, s'inscrit dans la continuité des travaux de modélisation en statistique spatiale, en lien avec des données réelles, notamment celles fortement liées au territoire de Mayotte. Présenté dans la dernière section, ce projet est prévu pour s'étaler sur le moyen et long terme.

7.1 Modèles stochastiques : Estimation, Approximation et Apprentissage automatique

Ce premier axe se positionne dans le prolongement de mes derniers travaux de recherche. Certains modèles probabilistes comme les équations différentielles stochastiques et les modèles (non) markoviens sont souvent utilisés en épidémiologie. La théorie de l'estimation des paramètres de ces modèles est encore ouverte à de nombreuses questions de

recherche. Le cas des équations différentielles stochastiques dirigées par des mouvements browniens ou des processus de Lévy généraux avec des moments d'ordre deux finis a été bien développée. Cependant dans le cadre des processus stables, la méthode classique du maximum de vraisemblance ne s'applique pas dans ce contexte car le rapport de vraisemblance n'existe pas. Une approche intéressante pour contourner ce problème est la méthode d'ajustement de trajectoire combinée à la technique des moindres carrés pondérés dans le cadre des processus d'Ornstein-Uhlenbeck pilotés par des processus stables. D'autres approches sont également considérées comme l'estimation non paramétrique. L'estimation statistique pour des modèles (non) markoviens, extension des Chaînes de Markov, est un sujet de recherche d'actualité et permet là aussi une analyse plus complète dans certains schémas d'observation de phénomènes réels. Ces aspects viennent prolonger une partie de mon intérêt scientifique pour les chaînes de Markov et les processus stables depuis ma thèse de Doctorat.

On considère l'équation différentielle stochastique (EDS) dirigée par un processus α -stable :

$$dX_t = b(X_t)dt + \sigma(X_{t-})dZ_t, \quad X_0 = \eta,$$

où la fonction $b : \mathbb{R} \rightarrow \mathbb{R}$ (appelée dérive), est une fonction mesurable inconnue, et $\sigma : \mathbb{R} \rightarrow \mathbb{R}_+$ est une autre fonction (inconnue) (qui est considérée comme le terme de nuisance). Le processus de Lévy $Z = \{Z_t, t \geq 0\}$ est un processus α -stable défini sur un espace de probabilité filtré et η peut être une variable aléatoire indépendante de Z ou une constante réelle. On suppose que le processus est observé en temps discrets $\{t_i = i\Delta_n, i = 0, 1, 2, \dots, n\}$, où Δ_n est la fréquence de temps sur laquelle les observations sont faites. Des travaux préliminaires ont été entamés dans [87] consacré à l'exploration de nouvelles méthodes d'estimation statistique via l'approximation des solutions d'EDS dirigées par les processus α stables [86]. Une idée actuelle en guise de perspective est l'étude des estimateurs récursifs et l'étude du choix de la fenêtre dans le cadre non paramétrique par l'approche basée sur les moments infinitésimaux en exploitant les moments d'ordre p dans le cas α -stable via les techniques de troncature sur les processus α stables.

L'existence de solutions positives et la mise en place de nouvelles propriétés d'ergodicité des solutions d'EDS dirigées par des processus α -stables serait un atout en vue d'application aux données réelles. Par ailleurs, le fait de s'intéresser à des estimateurs récursifs vient de l'application motivant ces travaux, à savoir l'estimation au fil du temps, des indicateurs associés à l'estimation des paramètres. L'avantage des estimateurs récursifs sur leur version non récursive est que leur mise à jour, d'un échantillon de taille n à un échantillon de taille $n + 1$, nécessite beaucoup moins de calculs. Cette propriété est particulièrement importante dans le cadre de l'estimation de la densité, étant donné que le nombre de points auxquels la fonction est estimée est généralement très grand. La première version récursive de l'estimateur de densité à noyau de Rosenblatt, et la plus célèbre, a été introduite dans la littérature et a été largement étudiée. Par exemple, l'estimateur récursif (proposé par [53]) de Nadaraya-Watson est donné par :

$$r_n(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} = \frac{m_n(x)}{f_n(x)}, \quad (7.1)$$

avec

$$m_n(x) = (1 - \gamma_n)m_{n-1}(x) + \frac{\gamma_n}{h_n^d} K\left(\frac{x - X_n}{h_n}\right) Y_n,$$

et

$$f_n(x) = (1 - \gamma_n)f_{n-1}(x) + \frac{\gamma_n}{h_n^d} K\left(\frac{x - X_n}{h_n}\right).$$

Soit $(w_n)_n$ une suite de réels positifs telle que $\sum_{n \geq 1} w_n = \infty$. Lorsque le pas de temps est égal à $w_n \left(\sum_{i=1}^n w_k\right)^{-1}$, l'estimateur r_n peut être réécrit comme suit :

$$r_n(x) = \frac{\left(1 - \frac{w_n}{\sum_{i=1}^n w_k}\right) r_{n-1}(x) f_{n-1}(x) + \frac{w_n}{\sum_{i=1}^n w_k} K\left(\frac{x - X_n}{h_n}\right) Y_n}{\left(1 - \frac{w_n}{\sum_{i=1}^n w_k}\right) f_{n-1}(x) + \frac{w_n}{\sum_{i=1}^n w_k} K\left(\frac{x - X_n}{h_n}\right)}.$$

L'objectif donc de ce projet de recherche est d'étendre cette classe d'estimateurs dans le cas des données fortement mélangeant issues des observations d'un processus X solution de l'EDS précédente. Cette propriété peut être généralisée si on suppose que l'on a deux bases de données dont la première est de cardinale $n_1 \leq n - 1$ et la seconde de cardinale $n - n_1$. On peut vérifier que :

$$\begin{aligned} m_n(x) &= \prod_{j=n_1+1}^n (1 - \gamma_j) m_{n_1}(x) + \sum_{k=n_1+1}^{n-1} \left[\prod_{j=k+1}^n (1 - \gamma_j) \right] \frac{\gamma_k}{h_k^d} K\left(\frac{x - X_k}{h_k}\right) Y_k \\ &+ \frac{\gamma_n}{h_n^d} K\left(\frac{x - X_n}{h_n}\right) Y_n \\ &= \alpha_1 m_{n_1}(x) + \sum_{k=n_1+1}^{n-1} \beta_k \frac{\gamma_k}{h_k^d} K\left(\frac{x - X_k}{h_k}\right) Y_k + \frac{\gamma_n}{h_n^d} K\left(\frac{x - X_n}{h_n}\right) Y_n, \end{aligned}$$

avec $\alpha_1 = \prod_{j=n_1+1}^n (1 - \gamma_j)$ et $\beta_k = \prod_{j=k+1}^{n-1} (1 - \gamma_j)$, $\forall k = n_1 + 1, \dots, n - 1$.

Se placer dans les contextes sus-mentionnés présentent un certain nombre de difficultés. Ainsi dans cet axe du projet, nous traiterons des estimateurs à noyau récursifs pour lesquels récursif signifie que l'estimateur peut être considéré comme une combinaison linéaire de deux estimateurs, le premier étant basé sur le premier échantillon de taille n_1 et le second basé sur le second échantillon de taille $n - n_1$. Nous établirons également des critères pour l'étude du choix de la fenêtre dans le cadre non paramétrique par l'approche basée sur les moments infinitésimaux en exploitant les moments d'ordre p des solutions dans un premier temps dans le cadre des processus α -stables à saut positifs qui offrent des avantages simplifiés d'approximation des intégrales stochastiques liées. Les aspects d'apprentissage machine seront aussi explorés.

Par ailleurs, concernant les propriétés de mélange, une première approche d'étude est en cours afin d'établir de nouvelles propriétés d'ergodicité (de mélange) en passant par un couplage quantitatif en distances de Wasserstein et/ou variation totale des solutions (de Markov) d'équations différentielles stochastiques dirigées par des processus stables.

7.2 Estimation des paramètres des lois stables par des statistiques des extrêmes

Ce projet de recherche est dans la continuité de mes travaux actuels basés sur l'inférence statistique et l'estimation par la théorie des valeurs extrêmes. Un constat général est que, les modèles Gaussiens souvent utilisés dans de nombreuses applications ne couvrent pas de nombreuses situations de modélisation. Par exemple, il est possible que les données ne soient pas symétriques, présentent une décroissance rapide au niveau de la queue de la distribution. La famille des lois stables généralisent ces considérations évoquées et ont des applications dans de nombreux domaines d'études, notamment en santé et environnement, car ils intègrent à la fois l'asymétrie et les queues lourdes. Il existe plusieurs moyens d'estimer les distributions stables comme la méthode des quantiles proposée par McCulloch, ou la méthode par régression linéaire de Koutrouvelis. Dans le cas unimodal, il est possible d'effectuer une estimation par maximum de vraisemblance dans certains cas. Mais en général, l'absence d'une formule explicite des densités de probabilité rend difficile l'estimation directe par maximum de vraisemblance. Cependant dans le cadre des données où la distribution est multimodale, il est possible d'estimer les paramètres d'une loi stable par la méthode de mélange. L'estimation des mélanges de lois en utilisant l'algorithme EM est une démarche statistique bien connue à partir des lois multinormales et étendue à d'autres familles de lois. Cet algorithme simplifie beaucoup les calculs, car il permet d'estimer les paramètres de chaque groupe séparément, permettant ainsi de modéliser plus facilement la covariance des observations à travers le temps. Par exemple si on se place dans la perspective de multiples sources de données de suivi des temps d'apparition de symptômes de la maladie à coronavirus 2019 (Covid-19), l'estimation de la distribution de probabilité de cette variable épidémiologique est un objectif important, encore plus important lorsqu'on est en présence d'évènements (rares, extrêmes) de faibles probabilités. Modéliser cette distribution de temps d'apparition comme un mélange de lois de probabilités incluant les lois classiques, lois à queues lourdes y compris les lois α -stable et les lois d'extrémum généralisées est possible tout en incluant les données présentant des queues lourdes type α -stable, basées sur les algorithmes EM et ses variantes stochastiques. Une perspective de recherche actuelle, va consister à chercher des estimateurs non paramétriques à noyaux en adaptant la sélection de la fenêtre dans le cas stable par l'analyse de l'erreur du moment d'ordre p , avec $p < \alpha$. Ou encore obtenir des approximations du maximum de vraisemblance des lois α -stables, en passant par des lois de Pareto. Cela permettrait de considérer des mélanges de lois α -stables avec d'autres familles de lois des distributions extrêmes (en effet, des familles d'estimateurs de l'indice de queue de distribution ont été explorées dans des travaux avec El Hadji Deme et mentionnés plus haut). Cette dernière perspective de recherche d'estimation incluant les distributions extrêmes peut être présenté brièvement comme suit. L'un des résultats fondamentaux de la théorie des valeurs extrêmes, concerne la distribution limite de la variable du maximum : la loi des valeurs extrêmes généralisées (GEV) caractérisé par un indice des valeurs extrêmes (IVE) $\gamma > 0$. L'IVE contrôle le comportement des queues de

distribution. Son estimation a été largement étudiée dans la littérature. L'estimateur le plus populaire dans le cas i.i.d est celui de Hill basé sur les différences entre les logarithmes des statistiques d'ordre :

$$\hat{\gamma}_n^H = \frac{1}{k} \sum_{j=1}^k \log X_{[n+1-i,n]} - \log X_{[n-k,n]}$$

où $X_{[k,n]}$ est la k -ième statistique d'ordre et $k = k(n)$ vérifie : $k \rightarrow +\infty$, $k/n \rightarrow 0$, $n \rightarrow +\infty$. Considérons une segmentation des données centrées d'un échantillon d'une loi α -stable $X = (X_1, \dots, X_n)$ de longueur n en L segments qui ne se chevauchent pas, chacun de longueur K ($n = LK$). Posons $\bar{Y}_l = \log(\max\{X_i\})$ avec $i \in 1, \dots, K$. On peut décrire des estimateurs des exposants α et β des lois stables en se basant sur ces dernières quantités, voir par exemple [107]. Rappelons que les lois stables $S(\alpha, \beta, \sigma, \omega)$ sont asymptotiquement de type Pareto :

$$\lim_{x \rightarrow +\infty} x^\alpha \mathbb{P}(X > x) = C_\alpha \frac{(1 + \beta)}{2} \sigma^\alpha,$$

où

$$C_\alpha = \left(2 \int_0^{+\infty} x^{-\alpha} \sin(x) dx\right)^{-1} = \frac{1}{\pi} \Gamma(\alpha) \sin\left(\frac{\pi\alpha}{2}\right) \text{ ou } \frac{1 - \alpha}{\Gamma(2 - \alpha)} \cos\left(\frac{\pi\alpha}{2}\right).$$

Posons $\bar{X} = \max\{X_i \text{ avec } i \in 1, \dots, K\}$. Notons que \bar{X} représente le maximum d'un grand nombre K d'échantillons indépendants. Pour étudier la vitesse à laquelle \bar{X} tend vers l'infini lorsque $K \rightarrow \infty$ nous considérons la variable $Z = K^{-1/\alpha} \bar{X}$ qui est asymptotiquement donnée par la distribution de Fréchet. Et en calculant la fonction de répartition de $Y = \log(\bar{X})$, on montre que

$$\mathbb{P}(Y \leq y) = \mathbb{P}(Z \leq e^y \cdot e^{-\frac{1}{\alpha} \log(K)}) \sim \exp\left(-e^{-\alpha\left(y - \frac{1}{\alpha} \log\left(K C_\alpha \frac{1+\beta}{2} \sigma^\alpha\right)\right)}\right) = \exp(-e^{-\alpha(y - \lambda_K)}),$$

avec $\lambda_K = \frac{1}{\alpha} \log\left(K C_\alpha \frac{1+\beta}{2} \sigma^\alpha\right)$. Ce qui fait que Y est asymptotiquement décrite par la distribution de Gumbel avec ces paramètres dépendant de K . La moyenne et la variance sont données par

$$\mathbb{E}(Y) = \lambda_K - \frac{c_e}{\alpha}, \quad \text{Var}(Y) = \frac{\pi^2}{6\alpha^2},$$

où c_e désigne la constante d'Euler. De la même manière en partant de la relation

$$\lim_{x \rightarrow +\infty} x^\alpha \mathbb{P}(X < -x) = C_\alpha \frac{(1 - \beta)}{2} \sigma^\alpha,$$

et en posant Y' égal à $\log(\underline{X})$ avec $\underline{X} = \max\{-X_i \text{ avec } i \in 1, \dots, K\}$. On montre que la loi de Y' est asymptotiquement décrite par la distribution de Gumbel dont la moyenne et la variance sont aussi données par

$$\mathbb{E}(Y') = \lambda'_K - \frac{c_e}{\alpha}, \quad \text{Var}(Y') = \frac{\pi^2}{6\alpha^2} = \text{Var}(Y), \quad \text{avec } \lambda'_K = \frac{1}{\alpha} \log\left(K C_\alpha \frac{1-\beta}{2} \sigma^\alpha\right).$$

Un résultat particulièrement simple est obtenu pour la différence des moyennes des deux queues comme suit :

$$\mathbb{E}(Y - Y') = \frac{1}{\alpha} \log\left(\frac{1 + \beta}{1 - \beta}\right) \quad \beta \neq 1, -1 \quad \text{and} \quad \alpha = \frac{\pi}{2\sqrt{3}} \left(\frac{1}{\text{Var}(Y) + \text{Var}(Y')}\right),$$

et on trouve

$$\beta = 1 - \frac{2}{1 + e^{\alpha(\mathbb{E}(Y-Y'))}}.$$

Il est donc possible d'obtenir des estimateurs par la méthode des moments pondérés en prenant en compte l'analyse de la différence des moyennes des queues de distribution des logarithmes des échantillons maximum et minimum apparaissant dans l'estimateur de Hill. Concernant le paramètre σ , on peut aussi travailler directement sur \bar{X} et \underline{X} en utilisant l'expression de la loi GEV évoquée précédemment. Des familles d'estimateurs consistents et robustes de l'indice de queue de distribution ont été étudiées dans la littérature pour des lois GPD. Soit F dans le domaine d'attraction de $GEV_{\xi_0, \mu_0, \sigma_0}$. On a donc :

$$\lim_{n \rightarrow \infty} n(1 - F(b_n x + a_n)) = \left[1 - \xi_0 \frac{x - \mu_0}{\sigma_0} \right]^{-1/\xi_0}$$

On sait aussi que les dépassements d'un seuil u suffisamment élevé peuvent être modélisés par une loi de Pareto généralisée $GPD_{\xi, \sigma}$. Comme $F \in \mathcal{D}(GEV_{\xi_0, \mu_0, \sigma_0})$, on a :

$$\begin{aligned} \lim_{n \rightarrow \infty} P\left(\frac{X-a_n}{b_n} > u+y \mid \frac{X-a_n}{b_n} > u\right) &= \lim_{n \rightarrow \infty} \frac{n\bar{F}(b_n(u+y)+a_n)}{n\bar{F}(b_n u+a_n)} \\ &= \frac{1+\xi_0 \frac{u+y-\mu_0}{\sigma_0}}{1+\xi_0 \frac{u-\mu_0}{\sigma_0}} \\ &= \left\{ 1 + \frac{\xi_0 \frac{y}{\sigma_0}}{1+\xi_0 \frac{u-\mu_0}{\sigma_0}} \right\}^{-1/\xi_0} \end{aligned}$$

Utilisant l'approche Pareto généralisée pour modéliser les dépassements de u par une loi $GPD_{\xi, \sigma}$, on a :

$$\begin{aligned} 1 - \lim_{n \rightarrow \infty} P\left(\frac{X-a_n}{b_n} > u+y \mid \frac{X-a_n}{b_n} > u\right) &= 1 - \lim_{n \rightarrow \infty} P\left(\frac{X-a_n}{b_n} \leq u+y \mid \frac{X-a_n}{b_n} > u\right) \\ &= 1 - \left[1 - \left(1 + \xi \frac{y}{\sigma} \right)^{-1/\xi} \right] \\ &= \left(1 + \xi \frac{y}{\sigma} \right)^{-1/\xi} \end{aligned}$$

On en déduit :

$$\begin{cases} \xi &= \xi_0 \\ \sigma &= \sigma_0 + \xi_0(u - \mu_0) \end{cases}$$

On peut donc exprimer les paramètres de la GEV en fonction de ceux de la GPD, ce qui conduira à résoudre le système suivant de trois équations à trois inconnues :

$$\begin{cases} \xi &= \xi_0 \\ \sigma &= \sigma_0 + \xi_0(u - \mu_0) \\ \lambda_u &= 1 - \exp \left\{ -\frac{1}{N} \left[1 + \xi_0 \left(\frac{u-\mu_0}{\sigma_0} \right) \right]^{-1/\xi} \right\} \end{cases}$$

Le principal défi de ce projet de recherche est d'identifier les techniques de calcul les plus appropriées pour concevoir des méthodes d'estimation capables de s'adapter à la taille et à la complexité des problèmes réels. Il s'agit d'incorporer des notions de quantiles, des algorithmes stochastiques adéquats et des méthodes de Monte Carlo, tout en maintenant une haute qualité d'estimation.

7.3 Estimation statistique, modèles spatiaux et Apprentissage Automatique avec des données réelles

Les modèles mécanistes en dynamique des populations, basés sur les réponses physiologiques au niveau individuel aux variables motrices de l'environnement, ont été proposés depuis les années 1970. Pour expliquer et identifier les oscillations d'abondance d'une population, des approches diverses et complémentaires existent : d'une part les données d'observation i.e. les suivis de terrain (capture-marquage-recapture, comptage), les expérimentations et des expertises et d'autre part, les méthodes/outils d'analyse, i.e. les études statistiques, les modèles mathématiques et de simulation et finalement les systèmes d'information géographique (SIG). Les modèles dynamiques permettent de prédire l'évolution dans le temps et l'espace des effectifs d'une population sous différents scénarios (caractéristiques de la population, de la zone géographique, de la gestion de la population, etc.) choisis et maîtrisés. Le cadre de modélisation présenté ici est une particularisation, pour un choix simplifié de fonctions biodémographiques, du cadre mathématique permettant d'obtenir la distribution des individus dans une population structurée en âge (physiologique) et en temps.

Le modèle de McKendrick-Von Foerster

Soit un temps t et un âge chronologique a donnés. Soit $u(t, x)$ la population de moustique au temps t et avec un âge dans $(x, x + dx)$ de sorte que la population totale pour un âge compris entre x_0 et x_1 est :

$$N(t, x_0, x_1) = \int_{x_0}^{x_1} u(t, x) dx.$$

À un temps fixé t , considérons la fraction de cette population composée d'individus d'âge x . Nous supposons que, passé un temps suffisamment petit h , ces individus meurent selon un pourcentage $\mu(t, x)h$ avec l'équation suivante :

$$u(t + h, x + h) = (1 - \mu(t, x)h)u(t, x),$$

et une fois divisé par h et en faisant tendre h vers 0, cela donne :

$$\partial_t u(t, x) + \partial_x u(t, x) = -\mu(t, x)u(t, x).$$

C'est ce qu'on appelle le modèle de McKendrick-Von Foerster. En supposant que la mortalité est nulle, ses termes peuvent être facilement interprétés : le flux (entrant) sortant d'individus vers la part de la population d'âge x à l'instant t est le même que le flux (sortant) entrant de cette part d'individus pour tout x et tout t . Une autre approche concerne l'étude des développements en fonction des sources d'influence environnementales, par le biais du concept d'âge physiologique. Ce développement peut être suivi au moyen d'un biomarqueur tel que la taille ou le stade, et peut être particulièrement intéressant lorsque nous travaillons avec des vecteurs dont le cycle de vie comprend plusieurs sous étapes distinctes et que nous voulions les suivre séparément. Soit $v(t, x)$ la vitesse de développement, que nous pouvons considérer comme la dérivée de x par rapport au temps t . Nous avons donc l'équation suivante :

$$\frac{dN(t, x, x + h)}{dt} = v(t, x)u(t, x) - v(t, x + h)u(t, x + h) - \int_x^{x+h} \mu(t, a)u(t, a)da.$$

Le terme $v(t, x)u(t, x)$ représente le flux entrant d'individus à la part de la population d'âges x et $x + h$ et le terme $v(t, x + h)u(t, x + h)$ le flux sortant. L'équation ci-dessus conduit à la version suivante de l'équation de Von Foerster en prenant en compte cette notion d'âge physiologique :

$$\partial_t u(t, x) + \partial_x(v(t, x)u(t, x)) = -\mu(t, x)u(t, x).$$

En ce qui concerne les conditions aux limites, Von Foerster a proposé que le flux entrant soit représenté par

$$v(t, 0)u(t, 0) = \int_0^{+\infty} \beta(t, a)u(t, a)da, \quad u(0, x) = u_0(x)$$

où β représente le terme de fécondité. Considérons le problème simplifié de l'estimation des paramètres v et μ dans le cas constant : $\partial_t u(t, x) + \partial_x(vu(t, x)) = -\mu u(t, x)$. On procède dans un premier temps par une méthode des différences finies afin de proposer un moyen de calculer une approximation numérique des valeurs :

$$u_j^{n+1} = u_j^n \left[\frac{1}{1 + \mu\Delta t} \left(1 - \frac{\Delta t}{\Delta x} v \right) \right] + \frac{1}{1 + \mu\Delta t} \frac{\Delta t}{\Delta x} v u_{j-1}^n; \quad j \in [[1; N_a]] \text{ et } n \in [[0; N-1]].$$

Si on introduit une source d'erreurs de lois gaussiennes (hypothèse) centrées ε_j^{n+1} dans l'approximation numérique on obtient :

$$u^{n+1} = u^n \left[\frac{1}{1 + \mu\Delta t} \left(1 - \frac{\Delta t}{\Delta a} v \right) \right] + u_j^n \left(\frac{1}{1 + \mu\Delta t} \frac{\Delta t}{\Delta a} v \right) + \varepsilon^{n+1},$$

avec

$$u^{n+1} = \sum_{j=0}^{N_a} u_j^{n+1}; \quad u^n = \sum_{j=0}^{N_a} u_j^n; \quad u_j^n = \sum_{j=1}^{N_a} u_{j-1}^n.$$

Ainsi par la méthode des moindres carrés ordinaires qui consiste à minimiser la quantité :

$$\sum_{n=0}^N (\varepsilon^{n+1})^2 = \sum_{n=0}^N \left[u^{n+1} - u^n \left[\frac{1}{1 + \mu\Delta t} \left(1 - \frac{\Delta t}{\Delta a} v \right) \right] - u_j^n \left(\frac{1}{1 + \mu\Delta t} \frac{\Delta t}{\Delta a} v \right) \right]^2,$$

on obtient les estimateurs (de vraisemblance) suivants :

$$\left\{ \begin{array}{l} \hat{\mu} = \frac{1}{\Delta t} \left(\frac{\sum_{n=0}^N (u^n)^2}{\left[\frac{\sum_{n=0}^N u^{n+1} u^n - \frac{\sum_{n=0}^N (u^n - u_j^n) u^{n+1} \sum_{n=0}^N (u^n - u_j^n) u^n}{\sum_{n=0}^N (u^n - u_j^n)^2} + \frac{\left(\sum_{n=0}^N (u^n - u_j^n) u^n \right)^2}{\sum_{n=0}^N (u^n - u_j^n)^2} \right]} - 1 \right) \\ \hat{v} = \frac{\sum_{n=0}^N \frac{\Delta t}{\Delta a (1 + \hat{\mu} \Delta t)} (u^n - u_j^n) \left(\frac{u^n}{1 + \hat{\mu} \Delta t} - u^{n+1} \right)}{\sum_{n=0}^N \frac{\Delta^2 t}{\Delta^2 a (1 + \hat{\mu} \Delta t)^2} (u^n - u_j^n)^2}. \end{array} \right.$$

De plus par discrétisation simple de $u(0, t) = \int_0^{N_a} \beta(x)u(t, x)dx$, on trouve : $u_0^{n+1} = \Delta a \sum_{j=0}^{N_a} \omega_j \beta_j u_j^n$, où les ω_j désignent des poids. L'idée de cette partie de mon projet de recherche est de contribuer à la mise en place de méthodes d'estimation en passant par la technique des moindres carrés non linéaires et d'établir ensuite des propriétés de consistance des estimateurs dans un cadre statistique. Les nouvelles approches en Apprentissage Automatique (Machine Learning) notamment l'approximation numérique par la méthode Physics Informed Neural Networks seront aussi explorées.

Modèles spatiaux, estimation de paramètres et apprentissage automatique

Les modèles statistiques spatio-temporels sont de plus en plus utilisés dans une grande variété de disciplines scientifiques pour décrire et prédire des processus spatialement explicites qui évoluent dans le temps. La prise en compte de la dépendance spatio-temporelle s'avère réaliste et nécessaire dans certaines situations. L'estimation statistique des paramètres dans de tels modèles est important pour la prédiction basée sur des données réelles. Soit $\{Y(\mathbf{s}, t) : \mathbf{s} \in D_s, t \in [[0, T]]\}$ un processus aléatoire spatio-temporel. Par exemple, $Y(\mathbf{s}, t)$ peut être le nombre de moustiques capturés dans l'environnement à une coordonnée géographique \mathbf{s} (latitude, longitude, profondeur) et à un instant donné t . Dans un contexte général d'un espace continu, on peut considérer un processus spatio-temporel du premier ordre donné par l'équation linéaire intégro-différentielle suivante (IDE) :

$$Y(\mathbf{s}, t) = \int_{D_s} W^\theta(\mathbf{s}, u)Y(u, t - \Delta)du + \eta(\mathbf{s}, t),$$

où $W(\mathbf{s}, u)$ est un noyau de transition qui peut dépendre d'un paramètre θ qui spécifie la "redistribution des poids" pour le processus à un temps précédent $t - \Delta$ sur le domaine spatial D_s , et $\eta(\cdot, t)$ est un processus spatial continu à moyenne nulle (pouvant varier dans le temps et statistiquement dépendant dans l'espace) de type gaussien de matrice de covariance Q et indépendant de $Y_{t-\Delta}$. En d'autres termes La valeur du processus qui nous intéresse Y en un lieu spatial donné s et à un instant donné t résulte de trois éléments : la valeur prise par ce que l'on appellera le noyau du modèle W^θ , la valeur du processus à l'instant précédent $t - \Delta$ et une erreur dite de capture modélisé par un bruit sous la forme d'un processus indépendant des valeurs prises. En particulier dans le cas d'un ensemble fini d'emplacements spatiaux de prédiction $D_s = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ le modèle s'écrit sous la forme ($\Delta = 1$)

$$Y_t(\mathbf{s}_i) = \sum_{j=1}^n W_{ij}^\theta Y_{t-1}(\mathbf{s}_j) + \eta(\mathbf{s}_i, t),$$

pour $t = 1, 2, \dots, T$ avec des poids de transition W_{ij}^θ dépendant à nouveau d'un paramètre θ qui spécifie la "redistribution des poids". En considérant le vecteur de valeurs $Y_t = (Y_t(\mathbf{s}_1), \dots, Y_t(\mathbf{s}_n))'$, l'équation ci-dessus peut être écrite sous forme de matrice vectorielle comme une autorégression linéaire du premier ordre :

$$Y_t = MY_{t-1} + \eta_t,$$

où la matrice de transitions $n \times n$ est donnée par M et le processus d'erreur spatiale additif $\eta_t = (\eta_t(\mathbf{s}_1), \dots, \eta_t(\mathbf{s}_n))'$ est indépendant de Y_{t-1} et est spécifié comme étant de moyenne

nulle et gaussien avec la matrice de covariance spatiale Q . Les dynamiques sont donc contrôlées par la matrice de propagation M et la covariance Q . Supposons que la suite du terme d'erreur η sont des erreurs de mesure iid de moyenne nulle qui sont indépendantes de Y . L'exigence d'une prévision non biaisée conduit à choisir θ de telle sorte que

$$\sum_{j=1}^n W^\theta(s_0, s_j) \hat{Y}_{s_j, t-\Delta} - Y(s_0; t) \sim 0.$$

De même

$$\begin{aligned} \text{Var}\left(\hat{Y}(s_0, t) - Y(s_0; t)\right) &= \text{Var}\left(\sum_{i=1}^n W^\theta(s_0, s_i) \hat{Y}(s_i, t - \Delta) + \sum_{i=1}^n W^\theta(s_0, s_i) \eta(s_i, t - \Delta) \right. \\ &\quad \left. - \sum_{j=1}^n W^\theta(s_0, s_j) Y(s_j, t - \Delta) - \eta(s_0, t)\right) \\ &= \text{Var}\left(\sum_{i=1}^n W^\theta(s_0, s_i) (\hat{Y}(s_i, t - \Delta) - Y(s_i, t - \Delta))\right) \\ &\quad + \text{Var}\left(\sum_{i=1}^n W^\theta(s_0, s_i) \eta(s_i, t - \Delta) - \eta(s_0, t)\right) \\ &= \mathbb{E}\left(\sum_{i=1}^n W^\theta(s_0, s_i) \eta(s_i, t - \Delta) - \eta(s_0, t)\right)^2. \end{aligned}$$

Posons

$$\lambda_i = W^\theta(s_0, s_i) \quad \delta(s_i; t - \Delta) = \eta(s_i, t - \Delta) \quad \delta(s_0; t) = \eta(s_0, t).$$

On a

$$\begin{aligned} \text{Var}\left(\hat{Y}(s_0, t) - Y(s_0; t)\right) &= - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \frac{1}{2} \text{Var}\left(\delta(s_i; t - \Delta) - \delta(s_0; t)\right) \\ &\quad + 2 \sum_{i=1}^n \lambda_i \frac{1}{2} \text{Var}\left(\delta(s_0; t) - \delta(s_i; t - \Delta) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \frac{1}{2} \gamma(s_i, s_j; t - \Delta, t)\right) \\ &\quad + 2 \sum_{i=1}^n \lambda_i \gamma(s_0, s_i; t - \Delta, t) = -\lambda^T \Sigma \lambda + 2\lambda^T \Sigma_0, \end{aligned}$$

où Σ est la matrice dont les éléments i et j sont les semi-variogrammes entre les points spatiaux s_i et s_j au temps $t - \Delta$. Le i -ième élément du vecteur Σ_0 est le semi-variogramme entre les points spatiaux s_i et s_0 au temps $t - \Delta$ et t : $\gamma(s_i, s_0; t - \Delta, t)$. L'estimation statistique associée à ces modèles spatio-temporels, y compris leurs versions non linéaires, est envisagée sur le long terme. L'objectif est de développer un cadre méthodologique, à la fois théorique et appliqué, qui pourrait convenir aux schémas d'observation des données temporelles et spatialisées disponibles, par exemple, sur les populations de moustiques. Des approches d'apprentissage automatique seront également explorées.

Liste de Publications



Parameter estimation for stable distributions and their mixture

Omar Hajjaji^a, Solym Mawaki Manou-Abi^{a,b} and Yousri Slaoui^a

^aLaboratoire de Mathématiques et Applications, Université de Poitiers, UMR CNRS 7348, Poitiers, France;

^bInstitut Montpellierain Alexander Grothendieck, Université de Montpellier, UMR CNRS 5149, Montpellier, France

ABSTRACT

In this paper, we consider estimating the parameters of univariate α -stable distributions and their mixtures. First, using a Gaussian kernel density distribution estimator, we propose an estimation method based on the characteristic function. The optimal bandwidth parameter was selected using a plug-in method. We highlight another estimation procedure for the Maximum Likelihood framework based on the False position algorithm to find a numerical root of the log-likelihood through the score functions. For mixtures of α -stable distributions, the EM algorithm and the Bayesian estimation method have been modified to propose an efficient and valuable tool for parameter estimation. The proposed methods can be generalised to multiple mixtures, although we have limited the mixture study to two components. A simulation study is carried out to evaluate the performance of our methods, which are then applied to real data. Our results appear to accurately estimate mixtures of α -stable distributions. Applications concern the estimation of the number of replicates in the Mayotte COVID-19 dataset and the distribution of the N-acetyltransferase activity of the Bechtel et al. data for a urinary caffeine metabolite implicated in carcinogens. We compare the proposed methods, together with a detailed discussion. We conclude with the limitations of this study, together with other forthcoming work and a future implementation of an R package or Python library for the proposed methods in data modelling.

ARTICLE HISTORY

Received 19 November 2023
Accepted 14 November 2024

KEYWORDS




Stable distribution;
parametric estimation;
Newton–Raphson algorithm;
bisection algorithm; mixture
model; EM algorithm; Gibbs
sampling algorithm;
Metropolis–Hastings
algorithm

MATHEMATICS SUBJECT CLASSIFICATIONS

97K80; 62-08; 62C05; 62G30;
62P10

1. Introduction

In recent decades, many researchers have shown an interest in studying α -stable distributions because of their ability to generalise widely used laws such as Gaussian, Lévy and Cauchy to handle impulsive and skewed data, which is particularly important in the financial field. In 1925, Lévy discovered that α -stable distributions arise as the limit of normalised sums of independent and identically distributed random variables. The family of α -stable distributions has skewness and tail thickness. Unfortunately, there is

CONTACT Solym Mawaki Manou-Abi  solym-mawaki.manou-abi@umontpellier.fr  Laboratoire de Mathématiques et Applications, Université de Poitiers, UMR CNRS 7348, Futuroscope Chasseneuil, Poitiers 86000, France;
 Institut Montpellierain Alexander Grothendieck, Université de Montpellier, UMR CNRS 5149, Place Eugene Bataillon, Montpellier 34090, France

no closed-form expression for the cumulative distribution function and the probability density function, except in a few cases such as the Gaussian, Lévy and Cauchy distributions. Let X be a α -stable random variable and $\phi(t) = E(\exp(itX))$ the characteristic function. Note that this family of laws has multiple parameterisations. We follow the presentation in [11,28,37,51] and consider two types of representations. A random variable $X \sim S(\alpha, \beta, \gamma, \zeta; 0)$ if $\phi(t)$ is expressed as follows

$$\phi(t) = \begin{cases} \exp\left(-\gamma^\alpha |t|^\alpha \left[1 + i\beta \left(\tan\left(\frac{\pi\alpha}{2}\right)\right) \text{sign}(t)(|t|^{1-\alpha} - 1)\right] + i\zeta t\right) & \text{if } \alpha \neq 1. \\ \exp\left(-\gamma |t| \left[1 + i\beta \frac{2}{\pi} \text{sign}(t) \log(\gamma |t|)\right] + i\zeta t\right) & \text{if } \alpha = 1. \end{cases} \quad (1)$$

where $\alpha \in (0, 2]$ is the index of stability that governs the heaviness of the tail, $\beta \in [-1, 1]$ the skewness parameter $\gamma > 0$ the scale parameter, $\zeta \in \mathbb{R}$ the location or shift parameter and the sign function defined by:

$$\text{sign}(t) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

A random variable X is $S(\alpha, \beta, \gamma, \mu; 1)$, [37] if,

$$\phi(t) = \begin{cases} \exp\left(-\gamma^\alpha |t|^\alpha \left[1 - i\beta \left(\tan\left(\frac{\pi\alpha}{2}\right)\right) \text{sign}(t)\right] + i\mu t\right) & \text{if } \alpha \neq 1. \\ \exp\left(-\gamma |t| \left[1 + i\beta \frac{2}{\pi} \text{sign}(t) \log(|t|)\right] + i\mu t\right) & \text{if } \alpha = 1. \end{cases} \quad (2)$$

These parameterisations are therefore usually characterised by four parameters: $\alpha \in [0, 2]$, $\beta \in [-1, 1]$, $\gamma > 0$ and ζ or μ . If $\beta = 0$ we have a symmetric stable distribution and the two parameterisations above are identical. If $\gamma = 1$ and $\mu = 0$ (or $\zeta = 0$), we say that X is a standard α -stable random variable. The different parameterisations have repeatedly led to misunderstandings. The parameterisation (2) proposed by Samorodnitsky *et al.* [37] does not provide continuity of the density function at points $\alpha = 1$ and $\beta = 0$ (because of the $\tan(\pi\alpha/2)$ term), nor does it provide a scale-location family at $\alpha = 1$ (because of the $\gamma \log(\gamma)$ term), while the parameterisation (1) is continuous with respect to all parameters. The above two formulations are connected by the key equation

$$\zeta = \begin{cases} \mu + \beta\gamma \tan\left(\frac{\pi\alpha}{2}\right) & \text{if } \alpha \neq 1 \\ \mu + \beta \frac{2}{\pi} \gamma \log(\gamma) & \text{if } \alpha = 1. \end{cases}$$

It is worth noting that α -stable distributions have infinite variance for any $\alpha < 2$, and the mean is infinite for $\alpha \in (0, 1]$, which makes estimating the parameters difficult. Note that the mean, if it exists ($\alpha > 1$), is the natural measure of the location, but it cannot be estimated as precisely as ζ . In the case of $\alpha = 2$, we get the Gaussian distribution, when $\alpha = 1$ and $\beta = 0$, it becomes the Cauchy distribution, and Lévy for $\alpha = \frac{1}{2}$ and $\beta = 1$. It is recommended to use the 0 parameterisation for numerical work and statistical inference. The standard mean-focus 1 parameterisation is suitable for studying algebraic properties, we

refer the reader to the work of Nolan [27]. Simulation methods are available for α -stable distributions, for example Kanter [15] was the first to generate α -stable random variables with $\alpha \in (0, 1)$. Later, Chambers, Mallows and Stuck [7] extended the method to the general case. It is important to note that when α is close to 1 or 0 with $\beta \neq 0$, the computations are typically much more numerically demanding and the results may not be very accurate, see `Matlab` [28] and `R` [50]. Since it is not known when exactly the numerical difficulties occur, we only warn the reader to pay great attention in their programs and practical purposes. In the `stabledist` package [50], the authors highlight and discuss such warnings. These warnings are not yet the subject of explicit studies, and in the future we plan to investigate the explicit numerical behaviour in the vicinity of these boundary points. The simulation of α -stable multivariate distributions has been the subject of research in [45]. The performance of our simulations is carried out with values of α that are not close to 1. In the context of real applications, our applications concern the case $\alpha \in (1, 2)$. For $\beta = 1$, the distributions are maximally skewed to the right, and the `FMStable` package [34] provides distribution functions that are faster and more accurate than the `stabledist` package.

There are several methods in the scientific literature to estimate the parameters of α -stable distributions, for example the Fractional Lower Order Moments (FLOM) method [19]; the quantile method improved by McCulloch [23], the Empirical Characteristic Function (ECF) and Maximum Likelihood (ML) method [28], although the lack of a closed-form expression for the probability density is a theoretical disadvantage. Finite mixture models are becoming increasingly popular and play a crucial role in density estimation, and the mixture of α -stable distributions is a popular tool for modelling skewed and impulsive data, making it applicable in various fields. Two well-known methods of mixture estimation are the Expectation–Maximisation (EM) algorithm and the Bayesian approach. The EM algorithm is particularly useful in cases where the data are incomplete or partially observed. Bayesian estimation uses prior knowledge of the probability distribution of the parameters to estimate the value of the unknown parameters. Much work has been done in this direction, including the papers cited by [6] and references therein. Let us mention the work of [44], which presents a stochastic EM algorithm for skewed α -stable distributions.

The main contributions of this paper to the literature are as follows. First, we introduce new estimators for parameters of the α -stable distribution, which mainly modify the well-known characteristic function estimator of this family by replacing the density function by a Gaussian kernel within the integral representation of the characteristic function. We also handle the estimation of the bandwidth parameter using the plug-in method proposed by [38] (for $\alpha > 1$) and [40,41] (for $\alpha < 1$). Second, the ML estimator is computed numerically using score functions and the False position algorithm. Furthermore, this allows to improve the EM algorithm of [6]. The Bayesian approach of [35] is improved by combining Gibbs sampling and the Metropolis–Hastings algorithm. We consider two-component models motivated by applications, but it can be generalised to finite-component univariate α -stable mixture models. The performance of the proposed estimators is compared with other candidates through a small-scale simulation study. All these methods have been used to estimate the reproduction number during the COVID-19 outbreak data in Mayotte. We also give application of a dataset related to the distribution of N-acetyltransferase activity data in the blood of 245 unrelated individuals for a caffeine urinary metabolites

involved in carcinogenic substances, which is available in [2] for two sub-populations study.

The paper is organised as follows. In Section 2, we first recall some related works such as the quantile-based methods originally developed by McCulloch [23] and introduce the kernel estimation method of the characteristic function and the estimation procedure for α -stable distributions. We also consider a new framework for the ML estimation of α -stable distributions based on the score function and the False position algorithm. Then, in Section 3, we outline an adapted Bayesian and EM algorithm approaches to estimate the parameters of a mixture of two α -stable distributions, which can be generalised to multiple mixtures. The performance of the above proposed estimators for estimating the parameters of α -stable distributions is given in Section 4. In Section 5, we apply the proposed methods with the above mentioned COVID-19 outbreak data in Mayotte together with the distribution of the above already mentioned N-acetyltransferase activity data for two sub-populations studies. We conclude with a section on the limitations of this paper and future theoretical and practical developments, in particular new stochastic approximation methods and algorithms.

2. Parameter estimation for single α -stable distributions

There are many approaches to estimate the parameters of single α -stable distributions.

2.1. Related works

Among the various approaches, including moment-based methods, the Empirical Characteristic Function (ECF), and the Maximum Likelihood (ML) method, we begin by briefly discussing a particular moment-based approach, namely the Fractional Lower Order Moments (FLOM) method, as described in [19], along with the improved quantile method proposed by McCulloch [23]. This method considers the lower-order fractional moments of the data to estimate the parameters of the distribution. It is also less sensitive to outliers and does not require any iterative optimisation procedures. However, a limitation of the FLOM method is that it assumes that the data are independently and identically distributed, which may not be the case in some real-world applications. Furthermore, as shown in [19], it does not guarantee a good estimate of the skewness parameter β . Another related approach is the quantile method, improved by McCulloch [23], which is based on the relationship between the q th quantile and the distribution parameters for any distribution. In the case of the α -stable distribution, the q th quantile depends on the scale and location parameters of the distribution as well as the α and β parameters, it is also robust to outliers and can handle missing data, but it may not work well for distributions with slowly varying tails or non-smooth density functions. In the same context, a new estimation algorithm for the tail index was proposed in [31] by considering a quantile conditional variance ratio. Based on the explicit formula (1) of the characteristic function, one can establish an estimation method using regression (ECF) [28]. Although the lack of a closed-form expression for the probability density is a theoretical disadvantage, note that in practice they are computed numerically using an integral transformation. It is in this way that many authors perform the ML method for parameter estimation [28]. The aim of this section is to present the relevant methods that we consider in order to improve our main results.

First, we recall the quantile method originally developed by McCulloch, as described in [23], provided by the `libstable4u` package [8]. We will use such a method to initialise our algorithms but with the constraint $\alpha \geq 0.6$ [12,23]. The method has been simplified in recent years, and the version we use is based on the work presented in [28], which can be described as follows. Let x_p be the p th quantile of the distribution $X \sim S(\alpha, \beta, \gamma, \zeta; 0)$ and define the following quantities:

$$\begin{aligned} v_\alpha(\alpha, \beta, \gamma, \zeta) &= \frac{x_{0.95} - x_{0.05}}{x_{0.75} - x_{0.25}}, & v_\beta(\alpha, \beta, \gamma, \zeta) &= \frac{x_{0.05} + x_{0.95} - 2x_{0.5}}{x_{0.95} - x_{0.05}}, \\ v_\gamma(\alpha, \beta, \gamma, \zeta) &= x_{0.75} - x_{0.25}, & v_\zeta(\alpha, \beta, \gamma, \zeta) &= -x_{0.5}. \end{aligned}$$

In [23], McCulloch provides tables from which the values of the above parameters can be derived, considering a standard α -stable distribution $Z \sim S(\alpha, \beta, 1, 0; 0)$. This is not restrictive, since the scaling property, as outlined in [37], allows to have $x_p = \gamma z_p + \zeta$, where z_p is the p th quantile of Z . The value of the above parameters could then be deduced from these quantities using the following relationships [28]:

$$\begin{aligned} v_\alpha(\alpha, \beta, \gamma, \zeta) &= v_\alpha(\alpha, \beta, 1, 0), & v_\beta(\alpha, \beta, \gamma, \zeta) &= v_\beta(\alpha, \beta, 1, 0), \\ v_\gamma(\alpha, \beta, \gamma, \zeta) &= \gamma v_\gamma(\alpha, \beta, 1, 0), & v_\zeta(\alpha, \beta, \gamma, \zeta) &= \gamma v_\zeta(\alpha, \beta, 1, 0) - \zeta. \end{aligned}$$

As we can see, v_α and v_β are independent of the scale and location parameters, and the above relationships allow one to obtain a reliable estimate of the four parameters when the sample set is large, see [28]. In the following, the vector parameter $\Theta_0 = (\alpha_0, \beta_0, \gamma_0, \zeta_0)$ will refer to the initial vector value parameter according to McCulloch’s method.

We now proceed to introduce the methods proposed in this work.

2.2. Characteristic function-based estimation method

Let $X \sim S(\alpha, \beta, \gamma, \zeta; 0)$ and let X_1, \dots, X_n be a sample of size n of X and F_n be the empirical cumulative distribution function:

$$\hat{\phi}_n^{(1)}(t) = \int_{\mathbb{R}} \exp(itx) dF_n(x) = \frac{1}{n} \sum_{j=1}^n \exp(itX_j).$$

Note that, given $t \in \mathbb{R}$, $\hat{\phi}_n^{(1)}(t)$ is a consistent estimator of $\phi(t)$ for large values of n (by a simple application of the law of large numbers). We introduce an alternative estimation of the characteristic function using the kernel method. We are concerned with the Gaussian kernel for the case when $\alpha > 1$, expressed as $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ for all $x \in \mathbb{R}$ and satisfying $\int_{\mathbb{R}} K(z) dz = 1$. This alternative estimate of the characteristic function is defined as follows:

$$\hat{\phi}_n^{(2)}(t) = \frac{1}{nh_n} \int_{\mathbb{R}} \exp(itx) \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right) dx,$$

where the bandwidth satisfies $\lim_{n \rightarrow \infty} h_n = 0$ in order to have the consistency. For practical purposes, and in the case of $\alpha > 1$, we choose the optimal bandwidth h_n by considering

the plug-in method proposed by Sheather and Jones [38], since it is based on the mean square error. On the other hand (specifically for $\alpha < 1$), we consider the second-generation plug-in method proposed by Slaoui in [40] (see also [41]). Set

$$\eta(\gamma t | \alpha; 0) = \frac{2}{\pi} t \log(\gamma |t|) \mathbb{I}_{\{\alpha=1\}} + \tan\left(\frac{\pi\alpha}{2}\right) \text{sign}(t) |\gamma|^{1-\alpha} (|t| - |t|^\alpha) \mathbb{I}_{\{\alpha \neq 1\}}. \quad (3)$$

We now define the following sample functions $g_n(t)$ and $h_n(t)$ to compute the estimated characteristic function at sample size n :

$$\begin{cases} g_n(t) = \Re(\hat{\phi}_n(t)), \\ h_n(t) = \Im(\hat{\phi}_n(t)), \end{cases}$$

where $\hat{\phi}_n(t)$ could be $\hat{\phi}_n^{(1)}(t)$ or $\hat{\phi}_n^{(2)}(t)$. We have:

$$\begin{cases} \sqrt{g_n(t)^2 + h_n(t)^2} = |\hat{\phi}_n(t)| = \exp(-\gamma^\alpha |t|^\alpha), \\ \arctan\left(\frac{h_n(t)}{g_n(t)}\right) = \arg(\hat{\phi}_n(t)) = -\gamma^\alpha \beta \eta(\gamma t | \alpha; 0) + \zeta t. \end{cases}$$

We also define the following quantities:

$$\begin{cases} y(t) = \log(-\log(|\phi(t)|)) = \log(\gamma^\alpha) + \alpha \log(|t|), \\ z_k = \arg(\hat{\phi}_n(t_k)), \\ B_k = \hat{\gamma}^{\hat{\alpha}} \eta(\hat{\gamma} t_k | \hat{\alpha}; 0). \end{cases}$$

The relationship between $y(t)$ and $\log(|t|)$ can be modelled linearly with a slope of α and an intercept of $a = \log(\gamma^\alpha)$. Set

$$y_k = \log(-\log(|\hat{\phi}_n(t_k)|)) = \log\left(-\log\left(\sqrt{g_n(t_k)^2 + h_n(t_k)^2}\right)\right),$$

where t_k is chosen on a grid of equally spaced points t_1, \dots, t_m using the sample data. We use the following weighted least squares method to minimise

$$\begin{aligned} S(a, \alpha) &= \sum_{k=1}^m W_k (y_k - a - \alpha \log(|t_k|))^2 \\ (\hat{a}, \hat{\alpha}) &= \arg \min_{(a, \alpha)} S(a, \alpha), \end{aligned}$$

where the weights $W_k = \frac{1}{\sigma_k^2}$ and σ_k^2 denotes the variance of the k th observation. We get the following estimators:

$$\begin{aligned} \hat{\alpha}_m &= \left(\sum_{k=1}^m W_k \log(|t_k|) y_k - \frac{\sum_{k=1}^m W_k y_k}{\sum_{k=1}^m W_k} \times \sum_{k=1}^m W_k \log(|t_k|) \right) \\ &\times \left(\sum_{k=1}^m W_k \log(|t_k|)^2 - \frac{\sum_{k=1}^m W_k \log(|t_k|)}{\sum_{k=1}^m W_k} \times \sum_{k=1}^m W_k \log(|t_k|) \right)^{-1}, \end{aligned}$$

$$\hat{a}_m = \frac{\sum_{k=1}^m W_j y_k - \hat{a}_m \sum_{k=1}^m W_k \log(|t_k|)}{\sum_{k=1}^m W_j},$$

$$\hat{\gamma}_m = \delta_0 \exp\left(\frac{\hat{a}_m}{\hat{\alpha}_m}\right).$$

Then from the following formula

$$z_k = \beta B_k + \zeta t_k,$$

again using the method of weighted least squares as described above, we obtain

$$\begin{aligned} \hat{\zeta}_m &= \gamma_0 \left(\sum_{k=1}^m W_k B_k z_k - \frac{\sum_{k=1}^m W_k t_k z_k}{\sum_{k=1}^m W_k t_k B_k} \times \sum_{k=1}^m W_k B_k^2 \right) \\ &\quad \times \left(\sum_{k=1}^m W_k B_k t_k - \frac{\sum_{k=1}^m W_k t_k^2}{\sum_{k=1}^m W_k t_k B_k} \times \sum_{k=1}^m W_k B_k^2 \right)^{-1} + \zeta_0, \\ \hat{\beta}_m &= \frac{\sum_{k=1}^m W_k t_k z_k - \hat{\zeta} \sum_{k=1}^m W_k t_k^2}{\sum_{k=1}^m W_k t_k B_k}. \end{aligned}$$

Since $\hat{\phi}_n$ is consistent, the consistency of the estimated parameters $\hat{a}_m, \hat{\beta}_m, \hat{\gamma}_m$ could be obtained using the classical regression estimation method. A simulation study is carried out to evaluate the performance of the proposed estimation. The choice of m was taken from [16], which suggests choosing points t_k in the interval $[0.1, 1]$. We set the parameters γ_0 and ζ_0 using the McCulloch method mentioned above [23].

2.3. Maximum likelihood approximation method

In the following lines, we present a method for estimating α -stable distributions within the framework of ML. Since the probability density function does not have a closed-form expression, the classical ML method is computationally difficult in this context, because the likelihood ratio does not exist explicitly. For this reason, we first use a numerical approximation of the density function of α -stable distributions, which provides accurate estimates. We introduce the following method based on score functions and the False position algorithm, which is an intuitive way to estimate the underline parameters of α -stable distributions. Suppose we are in parameterisation 1 and $\alpha > 1$. Set

$$\begin{aligned} g_d(x | \alpha, \beta) &= \int_0^\infty \cos(xr + \beta \eta(r, \alpha; 1)) r^{d-1} \exp(-r^\alpha) dr \mathbb{I}_{\{0 < d < \infty\}} \\ &\quad + \int_0^\infty [\cos(xr + \beta \eta(r, \alpha; 1)) - 1] r^{d-1} \exp(-r^\alpha) dr \mathbb{I}_{\{-2 \min(1, \alpha) < d \leq 0\}}, \tilde{g}_d(x | \alpha, \beta) \\ &= \int_0^\infty \sin(xr + \beta \eta(r, \alpha; 1)) r^{d-1} \exp(-r^\alpha) dr \mathbb{I}_{\{-\min(1, \alpha) < d < \infty\}} \end{aligned}$$

$$\begin{aligned}
& + \int_0^\infty [\sin(xr + \beta\eta(r, \alpha; 1)) - xr]r^{d-1} \exp(-r^\alpha) \, dr \mathbb{I}_{\{\alpha > 1, -\alpha < d \leq -1\}}, h_d(x | \alpha, \beta) \\
& = \int_0^\infty \cos(xr + \beta\eta(r, \alpha; 1)) \log(r)r^{d-1} \exp(-r^\alpha) \, dr, \tilde{h}_d(x | \alpha, \beta) \\
& = \int_0^\infty \sin(xr + \beta\eta(r, \alpha; 1)) \log(r)r^{d-1} \exp(-r^\alpha) \, dr,
\end{aligned}$$

for $x \in \mathbb{R}$ and $d \in \mathbb{N}$. Note that $\eta(r, \alpha; 1) = -\tan\left(\frac{\pi\alpha}{2}r^\alpha\right)$ should not be confused with the previous one in (3) from parameterisation 0. Recall the following theorem from [28].

Theorem 2.1 (Stable score function): *Let $\alpha \neq 1$. The univariate α -stable density in the 1-parameterisation is given by*

$$f(x | \alpha, \beta, \gamma, \zeta; 1) = \frac{1}{\pi\gamma} g_1\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right).$$

Then the score functions are given by

$$\begin{aligned}
\frac{\partial f}{\partial \alpha}(x | \alpha, \beta, \gamma, \zeta; 1) &= \frac{1}{\pi\gamma} \left[\frac{\pi\beta}{2 \cos\left(\frac{\pi\alpha}{2}\right)^2} \tilde{g}_{1+\alpha}\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right) \right. \\
&\quad \left. + \beta \tan\left(\frac{\pi\alpha}{2}\right) \tilde{h}_{1+\alpha}\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right) - h_{1+\alpha}\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right) \right], \\
\frac{\partial f}{\partial \beta}(x | \alpha, \beta, \gamma, \zeta; 1) &= \frac{\tan\left(\frac{\pi\alpha}{2}\right)}{\pi\gamma} \tilde{g}_{1+\alpha}\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right), \\
\frac{\partial f}{\partial \gamma}(x | \alpha, \beta, \gamma, \zeta; 1) &= -\frac{1}{\pi\gamma^2} g_1\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right) + \frac{x - \zeta}{\pi\gamma^3} \tilde{g}_2\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right), \\
\frac{\partial f}{\partial \zeta}(x | \alpha, \beta, \gamma, \zeta; 1) &= -\frac{1}{\pi\gamma^2} \tilde{g}_2\left(\frac{x - \zeta}{\gamma} | \alpha, \beta\right).
\end{aligned}$$

Many equations, including most of the more complicated ones, can only be solved by iterative numerical approximation. There are many root-finding algorithms that can be used to obtain approximations to such a given root. One of the most common is Newton's method or the secant method, but it may fail to find a root under certain circumstances, and it can be computationally expensive since it requires a computation of the derivatives of the function. Other methods are needed and a general class of methods are the two-point bracket methods. The False position or Bisection algorithm is one of these. The convergence rate of the bisection method could possibly be improved by using a different solution estimate. The False position algorithm runs as the first iteration of the bisection algorithm and, essentially, the root is approximated by replacing the actual function by a line segment on the bracketing interval and then using the classic double False position formula on that line segment, see [47].

Obviously, this method requires a good range for each parameter. For this reason we consider the initial intervals with a given margin to the initial values $\Theta_0 = (\alpha_0, \beta_0, \gamma_0, \zeta_0)$ based on the quantile method. The same procedure is used for the other three parameters.

In the following we will refer to this method as ML-Second. At this stage, however, one may be disappointed that there is no theoretical study of the existence of the root solutions of the above score functions under the False position method. This will be addressed as a perspective, but we will only look at the convergence numerically. Note that one can also use the `optim` command in R to directly maximise the log-likelihood function and obtain the estimated parameters. This is referred to as ML-first. The performance and consistency of the estimation is illustrated in a simulation study.

3. Parameter estimation for mixtures of α -stable distributions

Finite mixture models are becoming increasingly popular and play a crucial role in density estimation, and the mixture of α -stable distributions is a popular tool for modelling skewed and impulsive data, making it applicable in various fields. A well-known method for mixture estimation is the EM algorithm. We will consider the ECF (2.2) and ML (2.3) methods for estimating the four parameters of these distributions, in order to select the most efficient of them for use in the EM algorithm. This algorithm is particularly useful in cases where the data are incomplete or partially observed. Much work has been done in this direction, including the papers cited by Castillo-Barnes *et al.* [6] and references therein. We should also mention the work of Teimouri *et al.* [44,46], which presents a methodology for a stochastic EM algorithm applied to α -stable distributions. In [43] the author established a formula that includes symmetrical and asymmetrical α -stable distributions to estimate parameters using the EM algorithm, but the paper is still unpublished and we note that their algorithm does not converge for non-symmetric α -stable distributions. Since the Bayesian approach is also a useful and efficient tool for parameter estimation in mixture models, we have proposed a modified framework that involves updating the posterior distribution until it converges to the stationary distribution by combining Gibbs sampling and the Metropolis–Hastings algorithm, as described in [35]. The novelty of this latest approach, compared to [35], is essentially the choice of the rejection zones in the Metropolis–Hastings algorithm, which, in addition to updating the parameters, can significantly affect the estimates. Mixture models in general could be used to understand the development of an epidemic by estimating the generation time and the number of reproductions; see for example the work in [21], where the authors consider mixtures of Weibull, log-normal and other distributions to estimate the effective reproduction number during the COVID-19 outbreak on the island of Mayotte, France. Recall that the mixture of α -stable distributions requires careful initialisation and proper selection of its components. The K -means clustering [39] allows a better estimation of the components. Two common methods to infer parameters in mixture models are: the EM algorithm and the Bayesian approach, [6,35]. Bayesian estimation is a framework for formulating statistical inference problems. When predicting or estimating a random variable or process, the Bayesian philosophy is based on combining the evidence contained in the random variable with prior knowledge of the probability distribution of the random variable. The Bayesian estimation method uses prior data to estimate the value of the unknown parameters. This reduces the difference between the estimate and the true value of that parameter. In Bayesian modelling, the choice of priors then plays a crucial role in determining the posterior inference. The EM algorithm is a widely used computational method for estimating the parameters of statistical models with latent or missing variables. This algorithm is particularly useful in cases

where the data are incomplete or partially observed. The EM algorithm works by iteratively alternating between the E-step, where we estimate the expected value of the unobserved or latent variables given the current parameter estimates, and the M-step, where we maximise the likelihood of the observed data based on the expected values obtained in the E-step. This alternating process continues until convergence is achieved, resulting in the optimal parameter estimates for the model. Let n be the number of observations and z_i the latent observations with $i = 1, \dots, n$. Denote by $\lambda_1 = \mathbb{P}(i \in \{1, \dots, n\}, z_i = 1)$ the weight for the first component $j = 1$ and $\lambda_2 = 1 - \lambda_1$. In this paper we assume that we are in a two-component mixture. Of course, this can be generalised to more than two components.

3.1. The proposed Expectation–Maximisation algorithm

In this paper, we present an adapted Expectation–Maximisation algorithm by incorporating the above parameter estimation tools, namely the ML estimation method (by means of score functions and the False Position Algorithm) and the estimation by means of the Empirical or Kernel (ECF) function, both for the vector parameter Θ and for updating the E-step in the EM algorithm. Such adjustments are important when selecting the appropriate parameter estimation method in the EM algorithm. For simplicity, we consider two subpopulations, say 1 and 2, and compute the values of the latent vector z as follows. A sample value x_i is assigned to population 1 with probability (the probability that $z_i = 1$, conditional on the observed value of x_i), given by the key equation

$$\begin{aligned} p_i &= \mathbb{P}(z_i = 1 \mid \Theta) \\ &= \frac{\lambda_1 f(x_i \mid \alpha_1, \beta_1, \gamma_1, \zeta_1; 0)}{\lambda_1 f(x_i \mid \alpha_1, \beta_1, \gamma_1, \zeta_1; 0) + (1 - \lambda_1) f(x_i \mid \alpha_2, \beta_2, \gamma_2, \zeta_2; 0)}. \end{aligned} \quad (4)$$

These are the actual weights assigned to observation i when calculating the expected log likelihood in the EM algorithm. Our proposed EM algorithm is given as follows. The subscript ‘ t ’ indicates the t th iteration of the EM algorithm to obtain the increment at each ‘iteration’. We initialised $\lambda_1^{(0)}$ using the R software sample function noted `sample()` to draw a sample with attach specific probability to assign elements to subpopulation 1 or 2 according to the previous value in Equation (4). We refer also to Remark 1 for other initialisation options.

Note that, for a mixture model of two α -stable distributions, it is identified in most cases by specifying $\zeta_1 < \zeta_2$. The EM algorithm thus fixes on one of the modes depending on its initialisation. In the Bayesian approach, one could simply include this in the priors for the two location parameters. The E-step is used to estimate and update the allocation parameter λ_1 based on the values of the previous iteration, and once this is done, we can apply the usual estimation methods such as ML or ECF to the observations of each of the two components. In the M-step, we also maximise the conditional expectation for the well-chosen specific tolerance ϵ until convergence. Incorporating the ECF method into the above EM algorithm has been shown to be a valid approach for mixtures of α -stable distributions with suitable initialisation and specific tolerance ϵ . Explicit and analytical studies for convergence by the ECF approach in the EM algorithm will be addressed as a perspective work. Let us now turn to another interesting method known for its flexibility, namely the Bayesian estimation method.

Algorithm 1 EM algorithm for mixtures of α -stable distributions

- 1: Initialisation of the model, with selection of a specific tolerance ϵ .
- 2: **repeat**
- 3: **E-step:** We compute the n values of the vector z .
- 4: **for** $i = 1, \dots, n$ **do**
- 5: A sample value x_i is assigned to the population 1 (and so we set $z_i = 1$) with probability

$$p_i^{(t)} = \frac{\lambda_1^{(t)} f(x_i | \alpha_1^{(t)}, \beta_1^{(t)}, \gamma_1^{(t)}, \zeta_1^{(t)}; 0)}{\lambda_1^{(t)} f(x_i | \alpha_1^{(t)}, \beta_1^{(t)}, \gamma_1^{(t)}, \zeta_1^{(t)}; 0) + (1 - \lambda_1^{(t)}) f(x_i | \alpha_2^{(t)}, \beta_2^{(t)}, \gamma_2^{(t)}, \zeta_2^{(t)}; 0)}$$

- 6: **end for**
 - 7: $\lambda_1^{(t+1)} = \frac{1}{n} \{\#z_i = 1\}$. So we use the expected log likelihood or ECF to get the new parameters for instance from component 1: $\Theta_1^{(t+1)} = (\alpha_1^{(t+1)}, \beta_1^{(t+1)}, \gamma_1^{(t+1)}, \zeta_1^{(t+1)})$
 - 8: **M-step:** We then independently maximise the new parameters for each case, for instance from the component 1 in the expected log-likelihood, which is denoted by $Q^{(t+1)} = \sum_{i=1}^n \log(f(x_i | \alpha_1^{(t+1)}, \beta_1^{(t+1)}, \gamma_1^{(t+1)}, \zeta_1^{(t+1)}))$, and the total log likelihood is the mean of the two cases.
 - 9: **until** convergence has been achieved : $|Q^{(t+1)} - Q^{(t)}| < \epsilon$.
-

Remark 3.1: As an initialisation of our proposed algorithm, we could implement our algorithm with a sufficient number of iterations and then check the convergence by plotting the values of the estimated parameters obtained at each iteration versus the iterations. A more rigorous option, consisted in using the initialisation which maximise the likelihood.

3.2. The proposed Bayesian algorithm

As mentioned above, the Bayesian inference framework allows us to build a hierarchical model in which the unknown quantities are estimated via additional information (objective or not very informative prior or non-informative) or available data (informative priors) using Bayes' rule. The Bayesian inference has the advantage of providing credible intervals on the behaviour of the likelihood function, taking into account any given information on the parameters. The priors chosen for this model are as follows. We consider non-informative uniform priors for the exponent parameter α and the skewness parameter β on their supports as in [4,20]. Such choices are also discussed as being appropriate. An inverse gamma distribution with initial parameters (1, 1) is chosen for the dispersion γ , and a normal prior with parameters (0, 5) is chosen for the location parameter ζ , and these priors are conjugate priors in Bayesian inference for the mean and variance [33]. Furthermore, in line with other work in the literature on mixing problems [24], the Beta distribution (the familiar 2-way special case of the Dirichlet distribution) is used as a prior for binomial proportions in Bayesian analysis [29]. We will consider the beta distribution $Beta(1, 1)$ as a conjugate prior for the weights. Although it could be useful to use the information we know about the properties of the parameters of α -stable distributions (e.g. they are bounded and as α tends towards 2, β has less influence on asymmetry), the priors for $\alpha, \beta, \gamma, \zeta$ were

chosen inspired by Salas-Gonzalez *et al.* [35] and assumed to be independent of each other to ensure that our model remains free of any unwanted biases. Speaking generally about other choices of priors, we hope to develop future research work over time by considering different choices of priors, such as Jeffrey's priors (well defined for the parameters of mixtures of distributions), which are a challenging task for α -stable distributions since they are not available in closed form, especially when investigating the behaviour of the Fisher information matrix [11,22,25].

Note that computing the joint posterior distribution of the above parameters, given the data and priors, is often analytically intractable due to the lack of a closed-form expression for α -stable densities. To overcome this problem, we will use Markov Chain Monte Carlo (MCMC) methods, more precisely a combination of Gibbs sampling and Metropolis–Hastings algorithms (see [35]), as illustrated in the following steps. The novelty of this latest approach, compared to [35], is essentially the choice of rejection zones in the Metropolis–Hastings algorithm, which can significantly affect the estimates in addition to the updating of the parameters. The standard deviation σ of the normal distribution used to select candidates in the Metropolis–Hastings procedure is 0.1. Any given choice of hyper-parameters for the priors will be updated in the Metropolis–Hastings acceptance zone so that they are not significant for the speed of convergence with this approach.

3.2.1. The weight distribution

We assume that the prior distribution of the weights follows a $Beta(\zeta)$ distribution with initial parameters $\zeta = (1, 1)$. Since $\mathbb{P}(z_i = 1)$ is equal to λ_1 for $i = 1, \dots, N$, where N is the number of observations, the full conditional distribution for $\lambda = (\lambda_1, \lambda_2)$ is also a beta distribution, with updated parameters $\zeta_1 + n_1$ and $\zeta_2 + n_2$, where for example n_i is the frequency of observations assigned to component $i = 1, 2$. Thus, the updated distribution for the weights is $Beta(\lambda_1 + n_1, \lambda_2 + n_2)$.

3.2.2. Updating the vector parameter Θ using MCMC

In this step we consider the Metropolis–Hastings sampling method. We generate a candidate parameter $\Theta_j^{\text{new}} = (\alpha_j^{\text{new}}, \beta_j^{\text{new}}, \gamma_j^{\text{new}}, \zeta_j^{\text{new}})$, for example $j = 1$, from a proposal distribution $q(\cdot | \cdot)$, and it is accepted with probability $A_{\Theta_j^{\text{new}}}$, defined by:

$$A_{\Theta_j^{\text{new}}} = \min \left(1, \prod_{i=1, z_i=j}^N \frac{f(x_i | \alpha_j^{\text{new}}, \beta_j^{\text{new}}, \gamma_j^{\text{new}}, \zeta_j^{\text{new}}; 0)}{f(x_i | \alpha_j^{\text{old}}, \beta_j^{\text{old}}, \gamma_j^{\text{old}}, \zeta_j^{\text{old}}; 0)} \times \frac{p(\Theta_j^{\text{new}})q(\Theta_j^{\text{old}} | \Theta_j^{\text{new}})}{p(\Theta_j^{\text{old}})q(\Theta_j^{\text{new}} | \Theta_j^{\text{old}})} \right).$$

We also assume that the priors are independent. Then we get

$$p(\Theta_j) = p(\alpha_j)p(\beta_j)p(\gamma_j)p(\zeta_j).$$

In this paper we choose a normal distribution for $q(\cdot | \cdot)$. By symmetry we conclude that

$$q(\Theta_j^{\text{new}} | \Theta_j^{\text{old}}) = q(\Theta_j^{\text{old}} | \Theta_j^{\text{new}}).$$

Then $A_{\Theta_j^{\text{new}}}$ become:

$$\min \left(1, \prod_{i=1, z_i=j}^N \frac{f(x_i | \alpha_j^{\text{new}}, \beta_j^{\text{new}}, \delta_j^{\text{new}}, \omega_j^{\text{new}}; 0)}{f(x_i | \alpha_j^{\text{old}}, \beta_j^{\text{old}}, \delta_j^{\text{old}}, \omega_j^{\text{old}}; 0)} \times \frac{IG(\delta_j^{\text{new}} | \alpha_0, \beta_0)N(\omega_j^{\text{new}} | \epsilon, k)}{IG(\delta_j^{\text{old}} | \alpha_0, \beta_0)N(\omega_j^{\text{old}} | \epsilon, k)} \right). \tag{5}$$

Now we sample a uniform variable u in $[0, 1]$. If $A_{\Theta_j^{\text{new}}} > u$, we accept the new candidate variables, otherwise we keep those from the previous iteration. The fact that we consider a single rejection zone associated with the vector parameter Θ is possible because the priors are assumed to be independent. Thus, the Markov chain $\tilde{\Theta}_n = (\alpha_n, \beta_n, \gamma_n, \zeta_n)$, where n is the iteration index, is stationary, unlike [35], where the authors consider a multiple Markov chain for each parameter without taking advantage of independence.

3.2.3. Updating the allocation parameter

At each iteration, it is necessary to predict which subpopulation each observation belongs to. We do this by computing the conditional probability, for example for $j = 1$ as in (4), which is the probability that the observation x_i belongs to the component $j = 1$. Note that this method requires ordered steps to converge to the correct distribution, similar to the approach described in [35], where reversible Markov chain Monte Carlo was used to determine the number of components in the mixture model. However, unlike the approach in [35], we consider Equation (5) as the rejection zone in the Metropolis–Hastings step for all parameters, which seems to be numerically more accurate (considering the bias generated) than considering rejection zones for each parameter separately. We therefore summarise our adapted method in the following algorithm:

Algorithm 2 Bayesian algorithm for mixtures of α -stable distributions

Require: Initialisation of weight parameters.

Require: Number of iterations N and burn-in M .

- 1: **for** $t = 1, \dots, N$ **do**
 - 2: Obtain weights $\lambda = (\lambda_1, \lambda_2)$ by drawing samples from a symmetric beta distribution $\lambda \sim \text{Beta}(\xi_1 + n_1, \xi_2 + n_2)$ where n_1 is the frequency of observations assigned to the first component and n_2 to the second.
 - 3: Update the parameters of the proposal distribution $q(\cdot | \cdot) = N(\cdot | \theta, \sigma)$, setting θ to the value of the previous iteration and choosing a small value for σ (the standard deviation).
 - 4: Sample new candidates $\Theta_j^{\text{new}} = (\alpha_j^{\text{new}}, \beta_j^{\text{new}}, \gamma_j^{\text{new}}, \zeta_j^{\text{new}})$ from the proposal distribution $q(\cdot | \cdot) = N(\cdot | \theta, \sigma)$ for each component.
 - 5: Accept Θ_j^{new} according to Equation (5) and set $\Theta_j^t = \Theta_j^{\text{new}}$, otherwise set $\Theta_j^t = \Theta_j^{t-1}$.
 - 6: **for** each observation x_i **do**
 - 7: Obtain the allocation variable z_i using Equation (4).
 - 8: **end for**
 - 9: **end for**
 - 10: Compute the mean parameters: $\Theta_j = \frac{1}{N-M} \sum_{k=M}^N \Theta_j^{(k)}$.
-

In the Bayesian algorithm above, we compute the mean of the posterior rather than its median and confidence intervals. In fact, the parameter vector of the mixture of α -stable distributions in the Bayesian method is a convergent Markov chain. Taking the median as an approximation for the parameter vector leads to numerical problems and there is no guarantee that the chain will converge. To numerically overcome such problems, we used Monte Carlo methods to estimate the parameter vector, which proved its effectiveness in the simulation part with well-selected burn-in period M .

Now that we have established the main methodology of interest in this paper, the next section will focus on assessing the effectiveness of the approaches using simulated data and then applying it to real data.

4. Simulations

Our primary objective in this section is to evaluate the effectiveness of these approaches proposed in Sections 2, 3.1 and 3.2 by using simulated data. We first discuss the performance of our proposed estimator based on the ECF and ML methods for estimating the parameters of single α -stable distributions. Secondly, we consider the case of mixture of α -stable distributions with different parameters. Bold values indicate the closest estimates to the true value in their respective rows when accessing the simulation performance. To assess the precision of the simulations' performance, we calculate the Mean Square Error (MSE) over the entire set of vector parameters Θ :

$$\text{MSE} = \frac{1}{\dim(\Theta)} \sum_{i=1}^{\dim(\Theta)} \left(\hat{\Theta}_i - \Theta_i \right)^2.$$

4.1. The case of single α -stable distributions

We evaluate the performance through a simulation study of these methods and assess the effect of the observation size n with varying parameter values. We consider the Gaussian kernel in the ECF method (denoted by ECF-Kernel) and the ECF method with empirical function is denoted by ECF-Empirical, as presented above. In addition, we used the ML method (First and Second). It is worth noting that when using the ML-Second method, one observes that numerical convergence holds for the case $\alpha > 1$ as mentioned in the methodology. For the case $\alpha \leq 1$, we have not been able to show efficient results because we had problems with the convergence of the special functions g_d , h_d , \tilde{g}_d and \tilde{h}_d . This will be evaluated in more detail in further work.

The results of these simulations are presented through Tables 1–8. We find that the estimation methods demonstrated satisfactory performance.

4.2. The case of mixture of α -stable distributions

We consider here, the framework of mixture estimation of two α -stable distributions with different parameters. The aim is to evaluate the accuracy and efficiency of the methods described above, namely our adapted EM algorithm (including the use of ML-First, ML-Second, ECF-Kernel, ECF-Empirical) to update the parameters in the M-step, and the

Table 1. Comparison of parameter estimation methods – configuration 1.

Parameter	True value	<i>n</i>	ECF-Kernel	ECF-Empirical	ML-First	ML-Second
<i>α</i>	1.6	500	1.6356	1.6023	1.6078	1.6151
		750	1.6191	1.5484	1.5697	1.7641
		1000	1.6075	1.5519	1.5443	1.6894
<i>β</i>	−0.8	500	−0.7541	−0.7566	−0.6692	−0.3714
		750	−0.8375	−0.7668	−0.7389	−0.8081
		1000	−0.7696	−0.7337	−0.7781	−0.7756
<i>γ</i>	5	500	5.0124	4.8583	4.8791	5.0498
		750	5.2009	4.9608	4.9314	5.4934
		1000	5.1666	4.9888	4.9418	5.2904
<i>ζ</i>	12	500	11.8914	11.9563	11.8727	12.8558
		750	12.2187	12.2930	12.2241	12.8333
		1000	12.1193	12.1905	12.2968	12.5753

Note: Bold values indicate the closest estimates to the true value in their respective rows.

Table 2. Mean square error – configuration 1.

<i>n</i>	ECF-Kernel	ECF-Empirical	ML-First	ML-Second
500	0.0038	0.0059	0.0119	0.2296
750	0.0224	0.0227	0.0148	0.2412
1000	0.0107	0.0108	0.0237	0.1059

Note: Bold values indicate the closest estimates to the true value in their respective rows.

Table 3. Comparison of parameter estimation methods – configuration 2.

Parameter	True value	<i>n</i>	ECF-Kernel	ECF-Empirical	ML-First	ML-Second
<i>α</i>	1.4	500	1.3848	1.3243	1.3588	1.3605
		750	1.3973	1.3613	1.3649	1.4420
		1000	1.3329	1.2985	1.3210	1.3924
<i>β</i>	0.5	500	0.4304	0.4464	0.5150	0.5800
		750	0.5173	0.5651	0.5523	0.6135
		1000	0.5381	0.5542	0.5280	0.6106
<i>γ</i>	2	500	1.9819	1.8854	1.9495	1.9278
		750	2.1368	2.0573	2.0668	2.2048
		1000	2.0910	2.0162	2.0397	2.2395
<i>ζ</i>	−10	500	−9.9746	−10.0041	−10.0032	−8.4292
		750	−9.9876	−10.0300	−10.0021	−9.2277
		1000	−10.0807	−10.1048	−10.0705	−10.8141

Table 4. Mean square error – configuration 2.

<i>n</i>	ECF-Kernel	ECF-Empirical	ML-First	ML-Second
500	0.0015	0.0054	0.0011	0.6201
750	0.0047	0.0024	0.0021	0.1632
1000	0.0051	0.0061	0.0033	0.1831

Bayesian method. Let $\Theta_1 = (\alpha_1, \beta_1, \gamma_1, \zeta_1)$ and $\Theta_2 = (\alpha_2, \beta_2, \gamma_2, \zeta_2)$; the density of the mixture model is given by: $f(x, \Theta; 0) = \lambda_1 \times f(x, \Theta_1; 0) + (1 - \lambda_1) \times f(x, \Theta_2; 0)$.

After applying the last four methods, we again find that, all the estimation methods show satisfactory performance when $\alpha \in (1, 2)$, see Tables 9–14. The corresponding graphs to visualise the plots are presented in Figures 1–12. However, we have a limitation of the numerical performance for the case where α_1 or $\alpha_2 \in (0, 1)$. In this scenario

Table 5. Mean square error – configuration 3.

n	ECF-Kernel	ECF-Empirical	ML-First
500	0.0799	0.0495	0.0026
750	0.0156	0.0114	0.0032
1000	0.0102	0.0070	0.0109

Table 6. Comparison of parameter estimation methods – configuration 3.

Parameter	True value	n	ECF-Kernel	ECF-Empirical	ML-First
α	0.8	500	0.9831	0.8523	0.7717
		750	0.8601	0.7891	0.7730
		1000	0.8614	0.7502	0.7510
β	0.8	500	0.7755	0.7249	0.8436
		750	0.8903	0.9398	0.8166
		1000	0.6330	0.7373	0.8056
γ	3	500	3.3389	3.1960	3.0887
		750	3.2231	3.0899	3.0994
		1000	3.0905	2.8824	2.9579
ζ	-12	500	-11.5866	-11.6107	-11.9937
		750	-12.0331	-12.1348	-12.0439
		1000	-11.9651	-12.0894	-12.1991

Table 7. Comparison of parameter estimation methods – configuration 4.

Parameter	True value	n	ECF-Kernel	ECF-Empirical	ML-First
α	0.6	500	0.7295	0.6118	0.5932
		750	0.7136	0.5949	0.5852
		1000	0.6826	0.6447	0.5801
β	-0.5	500	-0.3862	-0.5280	-0.5449
		750	-0.3669	-0.4413	-0.4831
		1000	-0.4074	-0.4941	-0.4436
γ	4	500	4.7518	4.0385	4.0950
		750	4.3565	3.7984	3.8458
		1000	4.4246	4.2515	3.7425
ζ	5	500	4.1715	4.4790	4.8560
		750	4.5660	4.7277	5.1534
		1000	4.7741	4.9808	5.1215

Table 8. Mean square error – configuration 4.

n	ECF-Kernel	ECF-Empirical	ML-First
500	0.3203	0.0684	0.0079
750	0.0865	0.0295	0.0119
1000	0.0616	0.0164	0.0211

there are computational problems for the component assignment (instability) during the initialisation problem, which again confirms the importance of better initialisation and convergence. This point will be addressed in the future to solve this problem. Recall that our ML-approximation method requires $\alpha > 1$.

The above simulation performance shows numerically accurate parameter estimation for the mixture of α -stable distributions with a given two components, which can be extended to more than two components.

Table 9. Comparison of methods for estimating mixture model-configuration 0.

Parameter	True value	Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
α_1	1.2	1.2032	1.2099	1.2886	1.1689
β_1	0.5	0.5015	0.2977	0.4095	0.4906
γ_1	1	0.9988	1.0338	1.0253	0.9638
ζ_1	-4.25	-4.1939	-4.1146	-4.1474	-4.1906
λ_1	0.6	0.6037	0.622	0.599	0.601
α_2	1.2	1.1812	1.4926	1.1579	1.1434
β_2	-0.5	-0.6354	-0.6280	-0.3839	-0.5497
γ_2	0.5	0.5062	0.5155	0.5103	0.4887
ζ_2	0.3	0.3043	0.2694	0.2808	0.3157

Note: Bold values indicate the closest estimates to the true value in their respective rows.

Table 10. Comparison of methods for estimating mixture model-configuration 1.

Parameter	True value	Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
α_1	1.2	1.1206	1.1897	1.1804	1.1170
β_1	0.5	0.4741	0.3107	0.4041	0.5095
γ_1	1	0.9805	0.9959	1.0263	1.0309
ζ_1	-2.5	-2.4334	-2.3928	-2.3867	-2.4098
λ_1	0.6	0.6166	0.602	0.617	0.632
α_2	1.7	1.6958	1.7426	1.7412	1.7488
β_2	0.5	0.4871	0.9308	0.8949	0.9999
γ_2	0.8	0.7604	0.8348	0.7400	0.7011
ζ_2	3	2.9495	2.9409	2.9855	2.9730

Note: Bold values indicate the closest estimates to the true value in their respective rows.

Table 11. Comparison of methods for estimating mixture model-configuration 2.

Parameter	True value	Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
α_1	1.4	1.4115	1.4655	1.4116	1.4969
β_1	-0.5	-0.5134	-0.5944	-0.6479	-0.9999
γ_1	0.8	0.8259	0.8302	0.7766	0.7787
ζ_1	1.5	1.5957	1.6116	1.5991	1.6686
λ_1	0.5	0.6072	0.6190	0.6080	0.6060
α_2	1.8	1.7398	1.8292	1.8203	1.6979
β_2	0.5	0.3467	1	1	0.9999
γ_2	1.2	1.2296	1.1782	1.1437	1.0805
ζ_2	5.3	5.4898	5.4519	5.3541	5.4004

Note: Bold values indicate the closest estimates to the true value in their respective rows.

Table 12. Mean square error for estimating mixture model-configuration 0.

Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
0.0024	0.0410	0.0182	0.0047

5. Applications

Let us now propose an application of the methods described above to real data.

5.1. Data sets

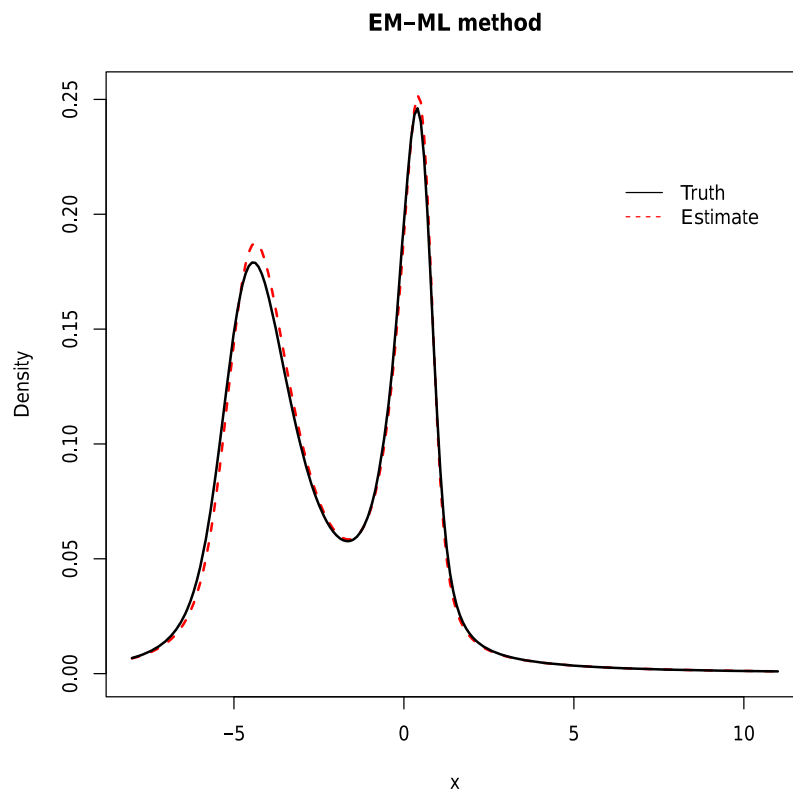
Let us recall the concept of serial interval (SI) in epidemiology. The serial interval refers to the time between the onset of symptoms in a primary case (the infector) and the onset of symptoms in a secondary case (the infectee). It plays a crucial role in understanding the

Table 13. Mean square error for estimating mixture model-configuration 1.

Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
0.0018	0.0266	0.0205	0.0310

Table 14. Mean square error for estimating mixture model-configuration 2.

Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
0.0095	0.0350	0.0333	0.0648

**Figure 1.** EM-ML-configuration 0.

dynamics of infectious disease transmission, as it helps estimate the potential speed and pattern of spread within a population. The generation time interval is then derived as the time interval from the infection of the infector to infection of the infectee. Thus, it is the time lag between infection in a primary case and a secondary case; and should be obtained from the time lag between all infectee/infector pairs [42]. As it cannot be observed directly, it is often replaced by the SI. Estimating the SI generation time and effective reproductive number [21] is an important task in understanding the development of an epidemic. In the previous paper [21] we only consider non-negative SI from the serial interval dataset of [14]. However, this dataset also contains negative serial intervals, because a suspected infector may show symptoms (infection) only after the infected person does. We also consider a data set related to the distribution of caffeine as a probe drug to determine the genetic status of two subpopulations of fast and slow acetylators. Acetylator status was determined

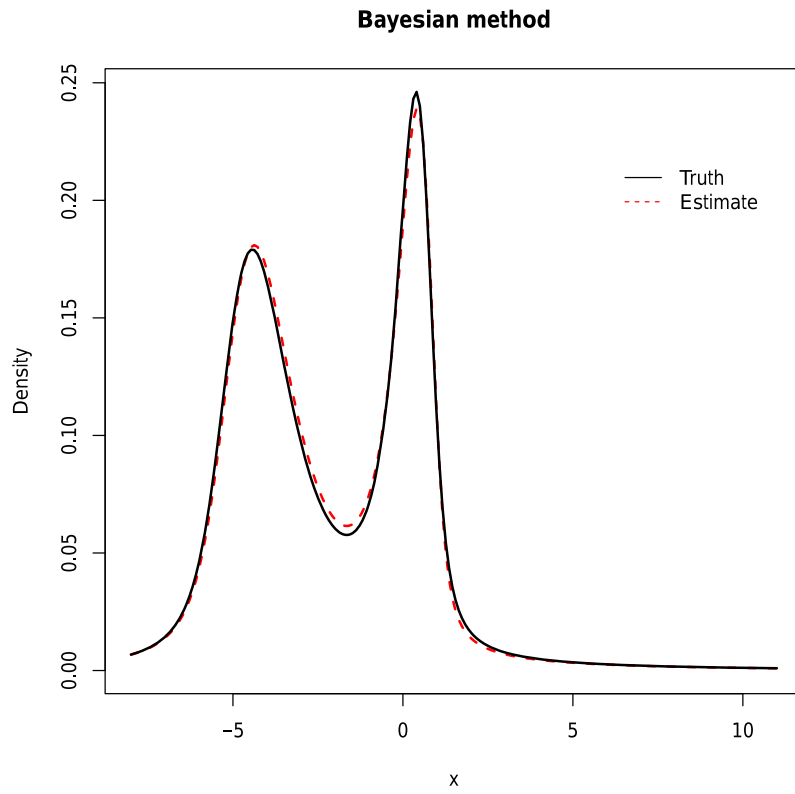


Figure 2. Bayesian-configuration 0.

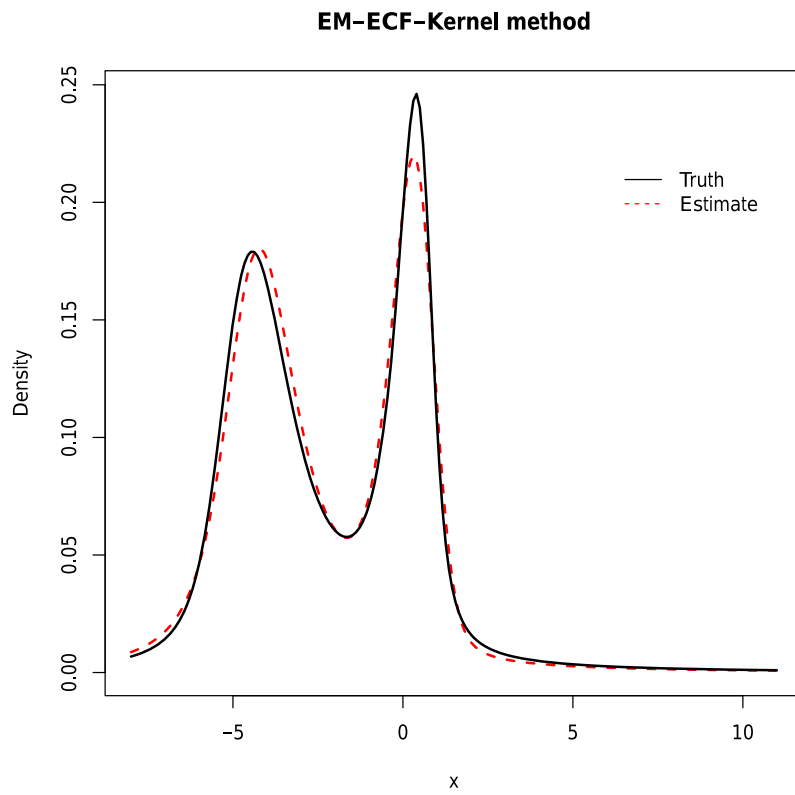


Figure 3. EM-ECF-Kernel-configuration 0.

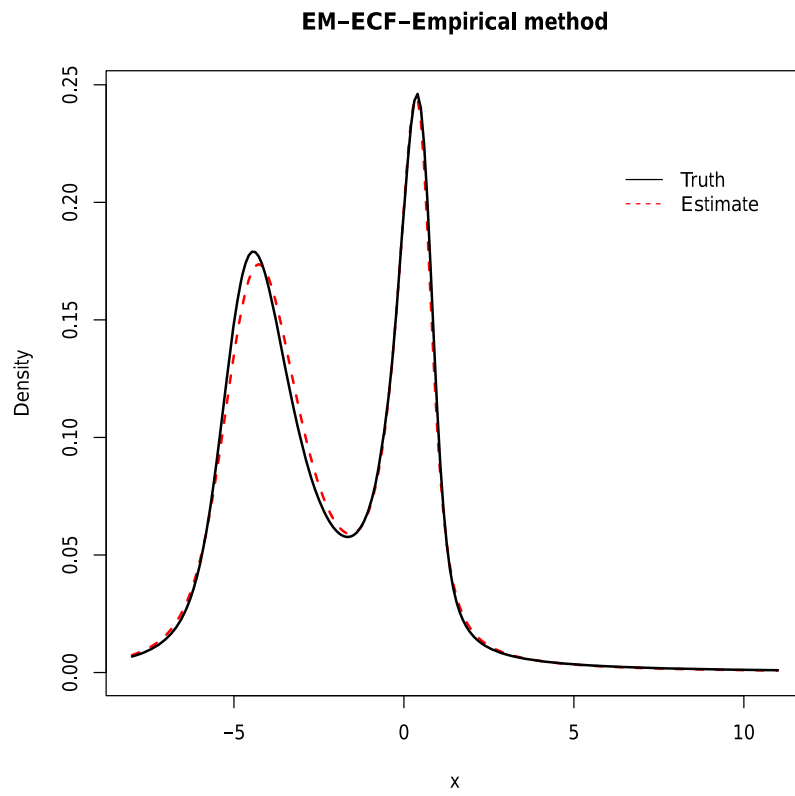


Figure 4. EM-ECF-Empirical-configuration 0.

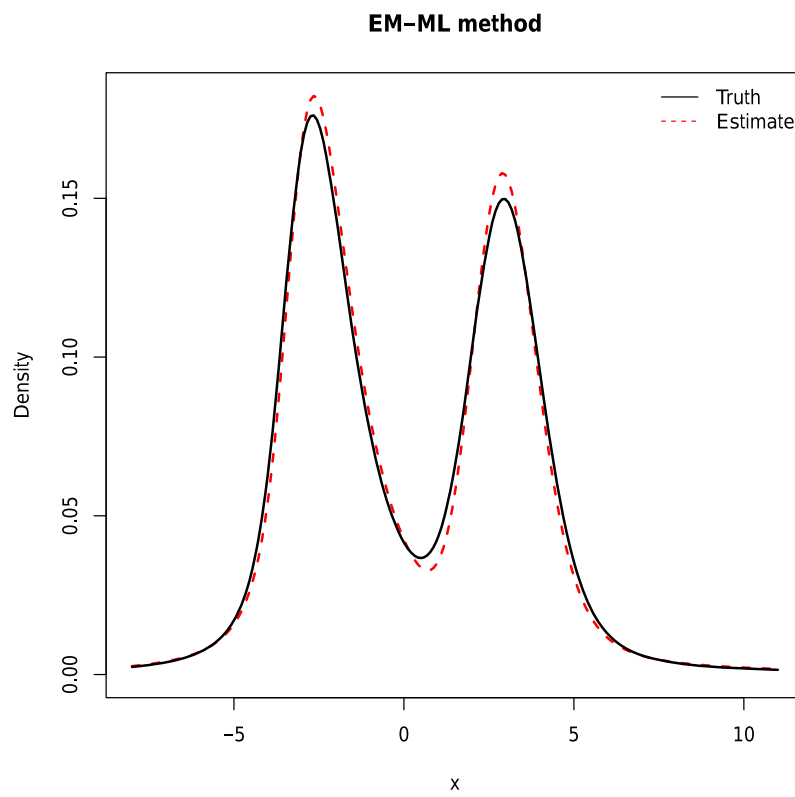


Figure 5. EM-ML-configuration 1.

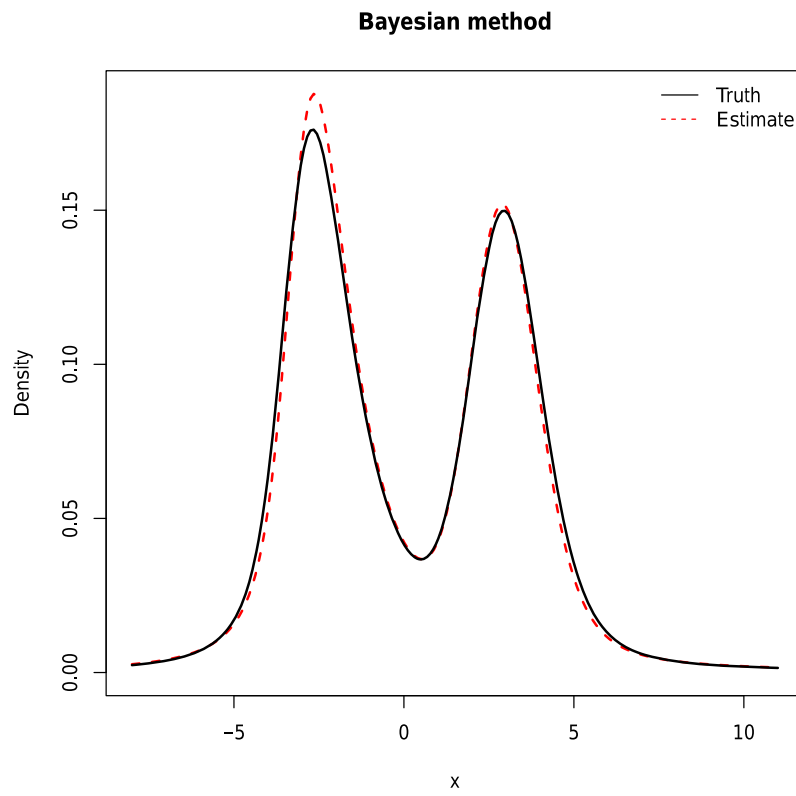


Figure 6. Bayesian-configuration 1.

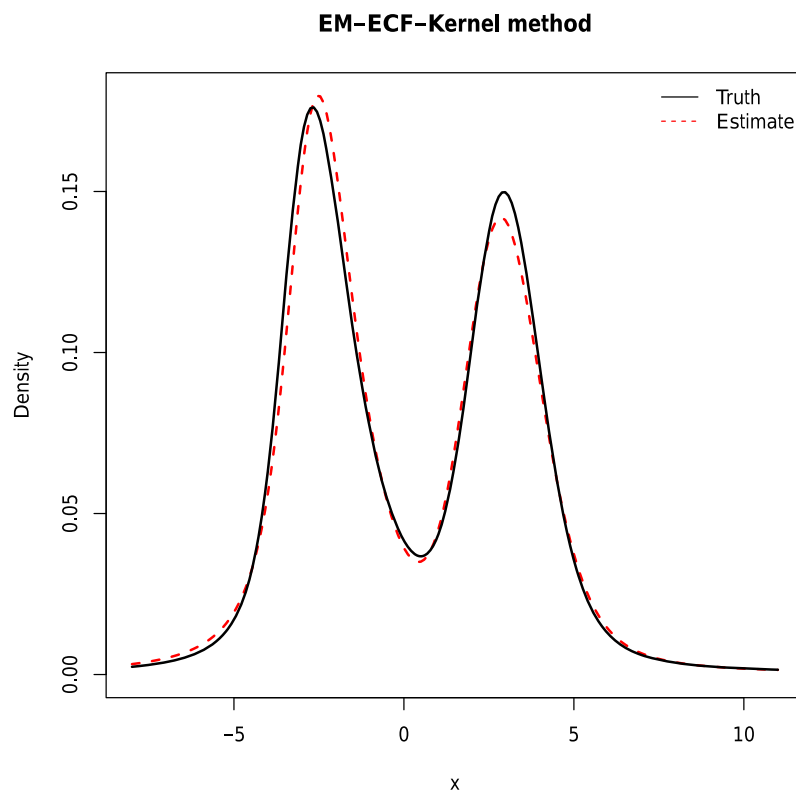


Figure 7. EM-ECF-Kernel-configuration 1.

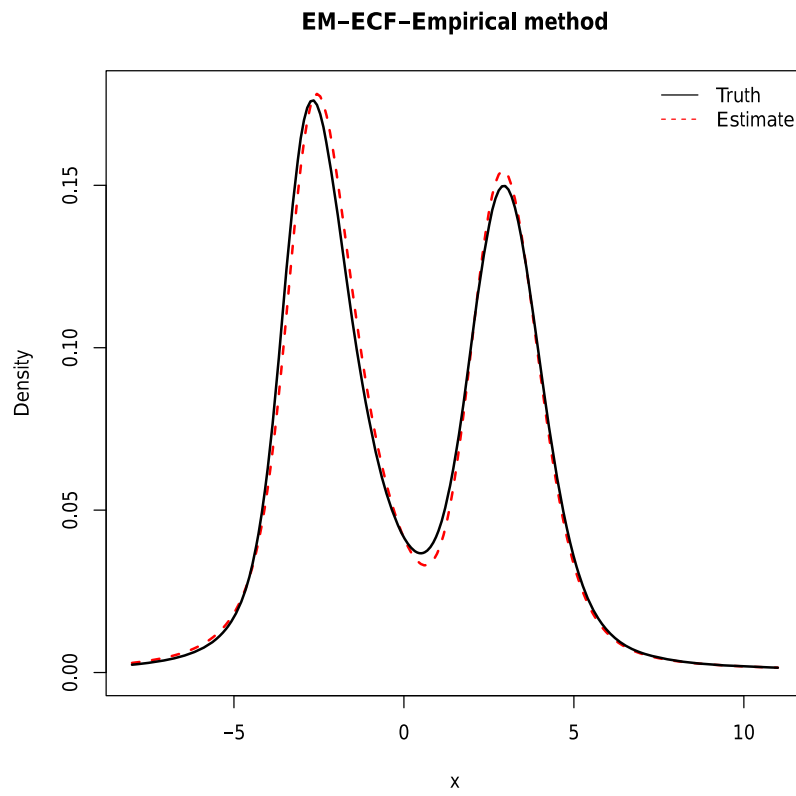


Figure 8. EM-ECF-Empirical-configuration 1.

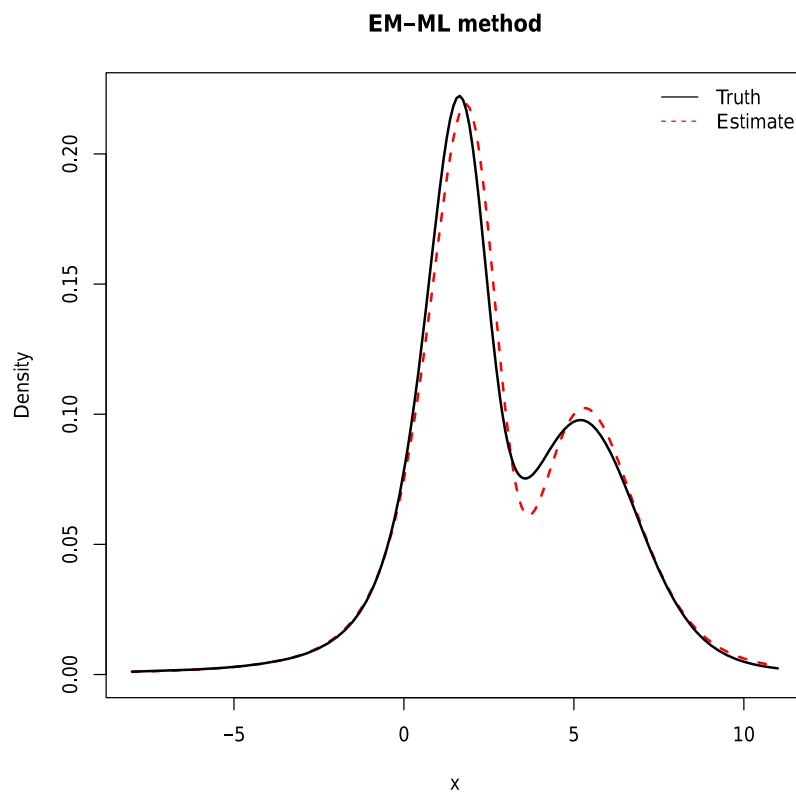


Figure 9. EM-ML-configuration 2.

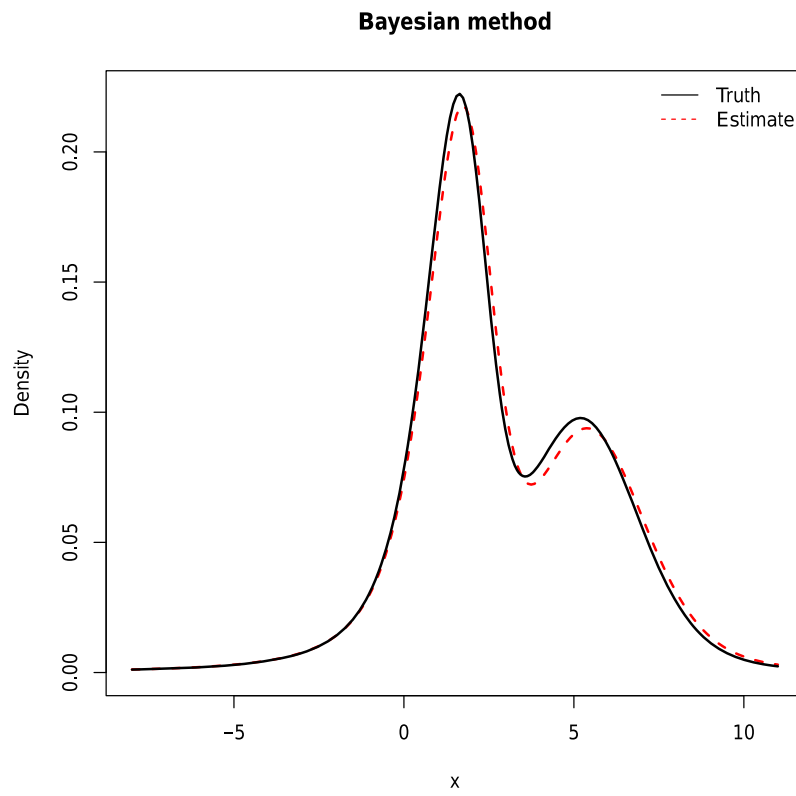


Figure 10. Bayesian-configuration 2.

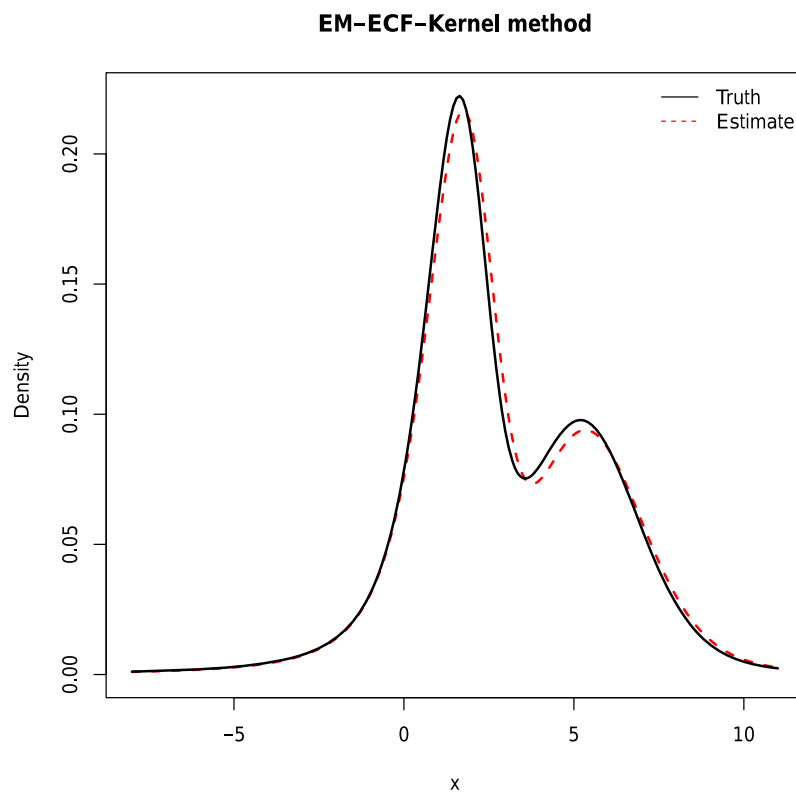


Figure 11. EM-ECF-Kernel-configuration 2.

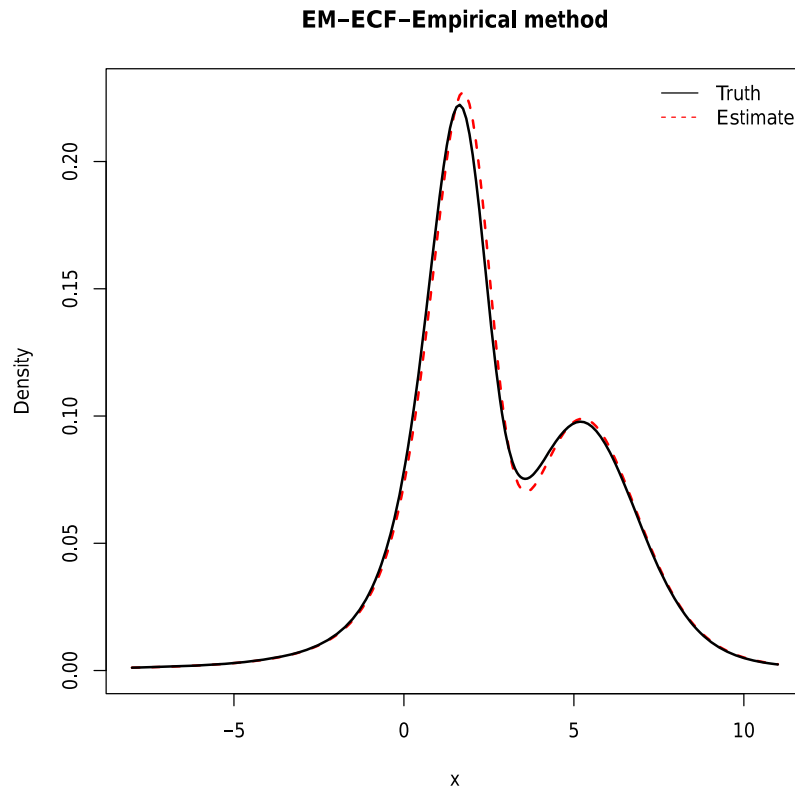


Figure 12. EM-ECF-Empirical-configuration 2.

from the urinary metabolic ratio for an enzymatic activity in the blood involved in the metabolism of carcinogenic substances, available in [2] for 245 unrelated individuals.

Mixture models have been applied to the distribution of the above Acetylator status dataset as it contains two subpopulations of slow and fast metabolisers within the population. In the case of the above SI dataset, although the histogram in Figure A2 shows a mode in the last bin around 18, the K -means method confirms the choice of a two-component mixture.

We perform a goodness-of-fit test to determine whether or not our sample data fits a normal distribution. As not all points lie approximately on the reference line in Figure A3, we can assume that the distribution of our data sets is not normally distributed (see also Table A1 for the Shapiro–Wilk, Anderson–Darling and Kolmogorov–Smirnov tests). In addition, we use skewness and kurtosis, which are two important measures in statistics; skewness refers to the lack of symmetry and kurtosis is a measure of whether or not a distribution is heavy-tailed. To calculate the skewness and kurtosis of our data sets, we use the `skewness()` and `kurtosis()` functions from the `moments` library in R software [17]. This allows us to see that our dataset distributions are skewed and not symmetric (see Table A2). Using the Kurtosis, we see that the distribution is leptokurtic showing heavy tails. We notice that the distributions of our data sets are skewed, not symmetric, and have heavy tails. Therefore, we propose to fit the above real data sets with α -stable (mixture) distributions. Note, however, that some statistical goodness-of-fit tests designed for α -stable distributions are available in the literature and require large sample sizes [1,5,18]. In particular, the recent work of [32] proposes a novel goodness-of-fit method based on

quantile (trimmed) conditional variances. Tables and comparisons of AIC and BIC criteria are provided to enhance understanding of the estimated parameters.

5.2. The serial interval and effective reproduction number in the context of Mayotte

The COVID-19 pandemic has caused considerable damage worldwide, disrupting productivity and causing widespread panic. In the French region of Mayotte, the regional health authority has made significant efforts to collect and monitor the spread of COVID-19, as documented in [21], by estimating the time-varying reproduction number, which is a non-pharmaceutical monitoring tool. The reported temporal daily cases of COVID-19 from 13 March 2020 to 11 January 2022 are shown in Figure A1.

By applying our methodology to the above serial interval data set, we found that the EM-ML method performed well in our analysis due to the small size of the data set, the peaked distribution and, the fact that the assignment vector changes at each iteration, which can affect the estimation of model parameters. This is also consistent with the work of [6]. To validate our decision, we calculated the AIC and BIC criteria for all four methods that led to this optimal choice. Although the AIC and BIC values were relatively close, we visually confirmed this choice in Figure 5. The tables associated with these methods are presented in Tables 15 and 16.

The estimated curves obtained by combining each method are shown in Figure 13. It can be seen that the mixture of two α -stable distributions using the EM-ML method provides the best fit of the serial interval.

The basic reproduction number R_0 at the start of an epidemic and the time varying (effective) reproduction numbers during an outbreak are important tools. Historically, it has been defined as the average number of new infections generated by an individual during a period of infectivity (see [10]). There are several methods to calculate this parameter. We consider a non-parametric approach [49] based on the generation time function associated with the serial interval distribution (see [3,30]). In the previous paper [21] we have only considered non-negative SI subset data from the [14] dataset and looked for SI estimation models such as Gamma, Lognormal and Weibull. Here we consider the entire dataset, including negative SI values, and deal with the α -stable mixture modelling framework. This is not a new dataset, but an original dataset from [14].

Let p be the probability distribution of the transmissibility of an infectious individual at age of infection τ , assuming that the entire population is susceptible. Let $\Gamma(t)$ be the number of new infections during the time interval $]t; t + dt[$. For discrete time $t \neq 0$, we have the following non-parametric formula for the effective reproduction number:

$$R_0(t) = \frac{\Gamma(t)}{\sum_{\tau \leq t} p(\tau)\Gamma(t - \tau)}. \tag{6}$$

Other improved methods of estimating the effective reproductive number exist in the literature, see [9,48] and references therein. In this work, using the epidemic incidence curve in Mayotte (between 13 March 2020 and 11 January 2022), we derive a generation time distribution to estimate the effective reproduction number $R_0(t)$ using the non-parametric

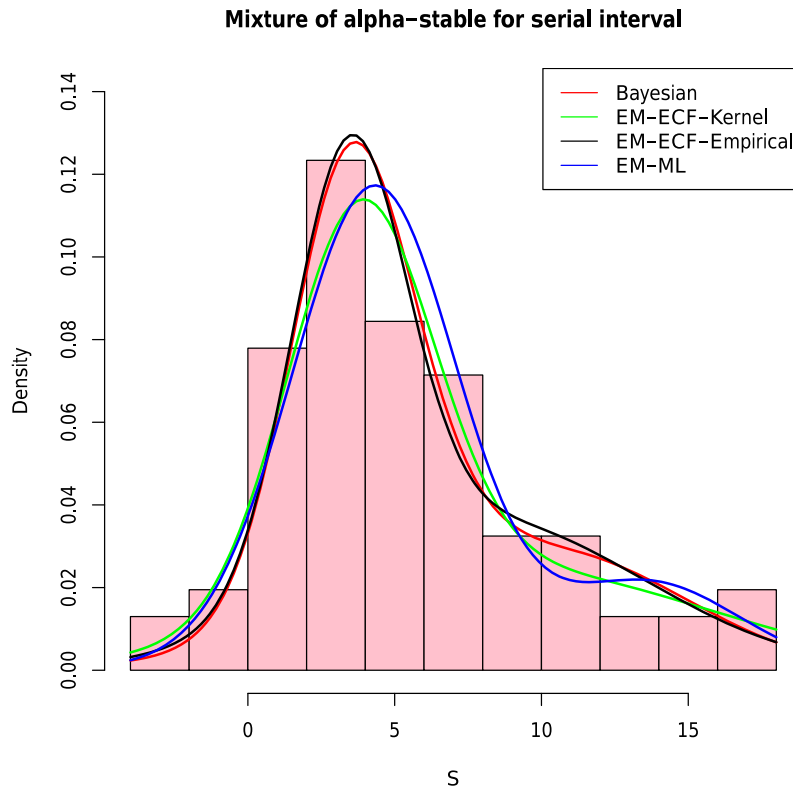


Figure 13. Mixture estimation of the serial interval distribution.

Table 15. Estimated parameters for the serial interval distribution.

Parameter	Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
α_1	1.7140	2	2	1.9614
β_1	0.4010	1	1	-0.8793
γ_1	1.6734	1.7524	1.3736	2.0038
ζ_1	3.6323	3.8096	3.4368	4.2990
μ_1	3.9558	3.8096	3.4368	4.2990
λ	0.7111	0.5974	0.5064	0.8311
$\mu = \lambda\mu_1 + (1 - \lambda)\mu_2$	5.7253	5.9183	5.7112	5.8660
α_2	1.7625	2	2	2
β_2	-0.6113	-1	-1	0
γ_2	3.2759	4.7511	3.8474	2.2267
ζ_2	10.8648	9.0474	8.0447	13.5769
μ_2	10.08101	9.0474	8.0447	13.5769

Table 16. Comparison table between the selection criteria for the serial interval distribution.

Methods	EM-ML	EM-ECF-Kernel	EM-ECF-Empirical	Bayesian
AIC	455.2616	457.9891	455.935	456.9538
BIC	476.3558	479.0833	477.0293	478.048

formula in Equation (6) and the best fit of the SI in a α -stable mixture modelling framework. The following plots in Figure 14 show the evolution of the time-varying reproduction number $R_0(t)$; we smooth the curve using estimated values.

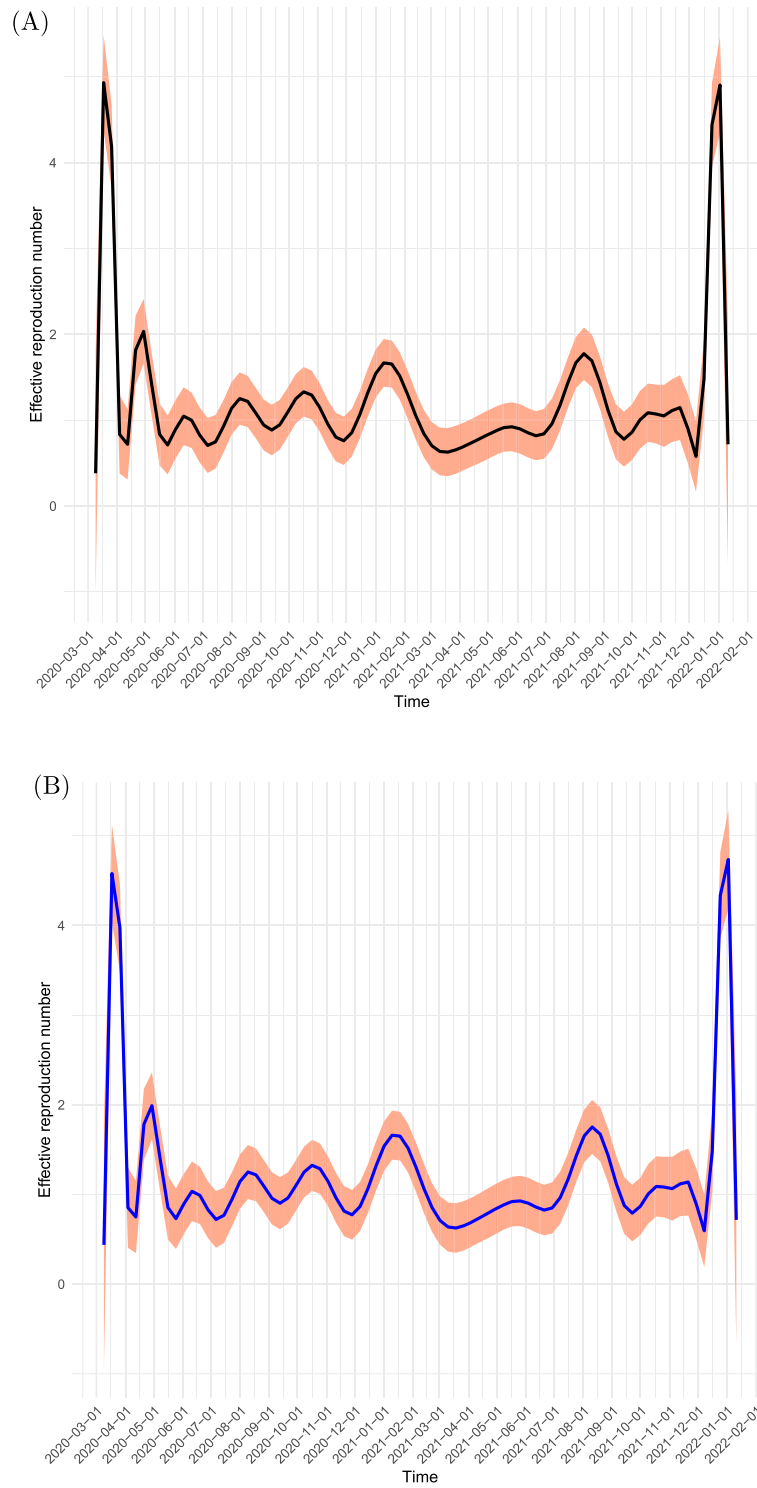


Figure 14. Evolution of the effective reproductive number in Mayotte from 13 March 2020 to 11 January 2022 with a mixture of α -stable distributions with a serial interval estimated by EM-ML (A) and estimated by EM-ECF-Empirical (B).

Table 17. Table of estimated parameters for the above distribution of N-acetyltransferase activity data.

Parameter	Bayesian	EM-ECF-Kernel	EM-ECF-Empirical	EM-ML
α_1	1.5936	1.9857	2	1.6939
β_1	0.8501	1	1	0.9999
γ_1	0.0549	0.0565	0.0499	0.0525
ζ_1	0.1756	0.1726	0.1741	0.1752
μ_1	0.2102	0.1738	0.1741	0.2025
λ	0.6188	0.5918	0.5755	0.6244
$\mu = \lambda\mu_1 + (1 - \lambda)\mu_2$	0.6273	0.5939	0.6059	0.6801
α_2	1.7175	1.8921	1.8864	1.3883
β_2	0.6134	1	1	0.9964
γ_2	0.2885	0.3409	0.3458	0.2273
ζ_2	1.2204	1.1447	1.1292	1.1502
μ_2	1.3045	1.2030	1.1915	1.4742

Table 18. Table comparing the selection criteria for the above distribution of N-acetyltransferase activity data.

Methods	EM-ML	EM-ECF-Kernel	EM-ECF-Empirical	Bayesian
AIC	109.2265	129.2364	129.3528	116.7802
BIC	140.7378	160.7478	160.8641	148.2916

If we look at the graphs of Figure 14, we see a complete similarity between them, and for both we can see that the effective reproduction number starts at a value of around 2.5, indicating that each infected individual infects, on average, 2–3 other people. Over time, we see a sharp increase in the reproduction number, peaking at around 4.5, indicating that the disease is spreading rapidly. This could be due to a number of factors, such as increased travel, relaxed social distancing measures, or a new variant of the disease that is more transmissible. After the peak, we see a decline in the number of reproductive cases, indicating that the disease is spreading more slowly. This could be due to interventions such as increased vaccination rates, stricter social distancing measures, or natural immunity acquired by those who have recovered from the disease. The reproduction number will eventually fall below 1, indicating that the disease is no longer spreading and can be considered to be under control. Eventually, however, the reproduction number will suddenly increase due to the effect of the new omicron variant.

5.3. The distribution of N-acetyltransferase activity data

Here we consider a dataset relating to the distribution of N-acetyltransferase activity in the blood of $n = 245$ unrelated individuals for a caffeine urine metabolite test for an enzyme involved in the metabolism of carcinogens available in Bethtel *et al.* [2]. This data set has been used in the past to test mixture models as it contains two subpopulations of slow and fast metabolisers within the population.

We observe that almost all the estimated parameters are similar, except for the stability index α , see Table 17. The EM-ML model is the most effective, as shown in Table 18, with the lowest AIC and BIC values. In addition, the corresponding figure for the four methods is shown in Figure 15.

The EM-ML method provides a better fit to the distribution of N-acetyltransferase activity in the blood data than the other approaches. This is because the other methods,

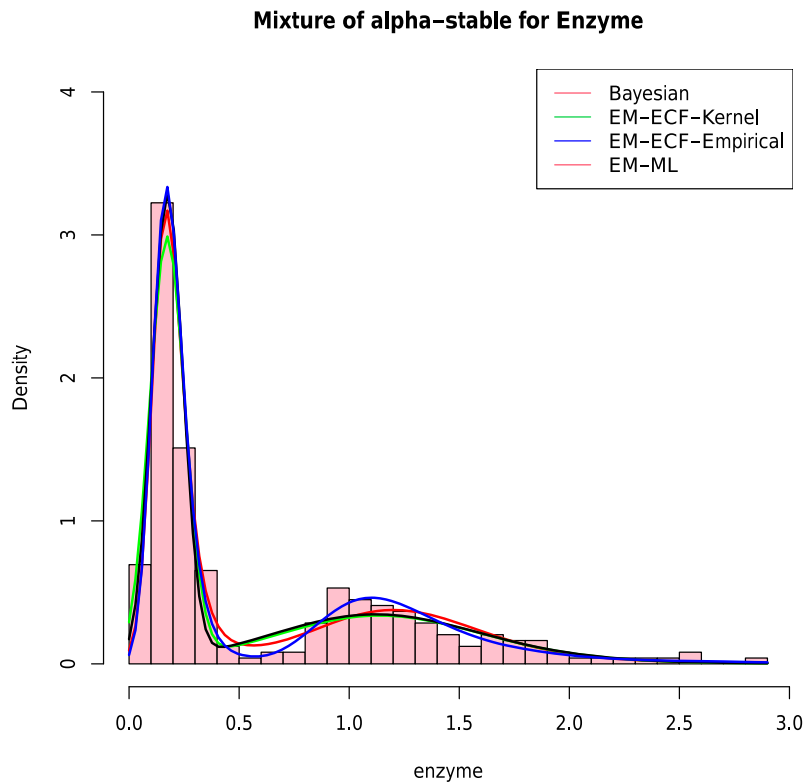


Figure 15. Mixture methods applied to the distribution of N-acetyltransferase activity data in the blood of 245 unrelated individuals in Bethtel *et al.* [2].

particularly EM-ECF, are sensitive to changes in the assignment vector at each iteration. Furthermore, the Bayesian approach still has the advantage of using any prior information on the parameters, which explains its efficiency on simulated data. However, it also requires certain conditions, such as the length of the observations, to ensure accurate results.

6. Conclusion and perspectives

In this paper, we consider the parameter estimation of univariate α -stable distributions and their mixture. We introduce some new techniques, such as the Gaussian kernel estimator in the characteristic function for the case $\alpha > 1$, which has shown more efficient performance than the empirical characteristic function in the simulation study. We also perform another estimation procedure in the ML framework based on the False position algorithm method to find the root of the log-likelihood through the score functions established in [28].

In the case of estimating the mixture of α -stable distributions, although we have limited our analysis to two components, the proposed methods can be generalised to multiple components.

The EM algorithm was adapted to estimate the parameters of each subpopulation to convergence by combining the ECF and ML methods in the M-step. Integrating the ECF into the M stage has proved to be a valuable practical approach.

The Bayesian method, which is more flexible but requires many steps to perform the estimation, has also been adapted to the parameter estimation of a given mixture model

of α -stable distributions. Note that there is a limitation to the initialisation of the above algorithms.

Finally, we consider two types of applications of our estimation methods on real data, namely the estimation of the reproduction number of the COVID-19 in Mayotte and the Acetylator status dataset.

In terms of future research, an immediate practical study is to address the question of the initialisation, followed by the implementation of an R package or Python library to further extend the use of such useful tools in data modeling for parameter estimation of a mixture of α -stable distributions as well as a focus on statistical goodness-of-fit tests designed for mixtures of α -stable distributions. We also plan to implement a more efficient approach to overcome the long burn-in period in the MCMC method using importance sampling [13]. Again, a future research direction would be to extend our study to the case of non symmetric α -stable distributions using Gibbs sampling [36].

We also aim to study some identity representations for α -stable distributions that incorporate a Weibull location scale mixture model, as stated in [43] in the symmetric case. In this way, our goal is to develop a stochastic algorithm that can account for the additional computational complexity [26] of the Weibull distribution.

Supplementary information

Additional supporting information on the original papers presented in the study and the R codes including datasets are available from the corresponding author on reasonable request.

Acknowledgements

The authors thank the reviewers whose comments and suggestions improved the paper. This work was funded by a research contract of the corresponding author with the Regional Health Agency (ARS) of Mayotte. We would also like to thank Julien Balicchi of the ARS of Mayotte for his contributions to data collection.

Declarations

The authors declare that they have no competing interests. The co-authors have seen and agree with the content of the manuscript. We certify that the submitted work is original and is not under consideration for publication elsewhere.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- [1] M.C. Beaulieu, J.M. Dufour, and L. Khalaf, *Exact confidence sets and goodness-of-fit methods for stable distributions*, J. Econom. 181 (2014), pp. 3–14.
- [2] Y.C. Bechtel, C. Bonaiti-Pellie, N. Poisson, J. Magnette, and P.R. Bechtel, *A population and family study N-acetyltransferase using caffeine urinary metabolites*, Clin. Pharmacol. Ther. 54 (1993), pp. 134–141.
- [3] P.Y. Boelle and T. Obadia, *R0: Estimation of R0 and real-time reproduction number from epidemics*, R package version 1.2-6. 2015; software available at <https://github.com/tobadia/R0>.
- [4] D. Buckle, *Bayesian inference for stable distributions*, J. Am. Stat. Assoc. 90 (1995), pp. 605–613.

- [5] K. Burnecki, A. Wylomanska, and A. Chechkin, *Discriminating between light-and heavy-tailed distributions with limit theorem*, PLoS. ONE. 10 (2015), pp. e0145604.
- [6] D. Castillo-Barnes, F.J. Martínez-Murcia, J. Ramírez, J. Górriz, and D. Salas-Gonzalez, *Expectation-maximization algorithm for finite mixture of α -stable distributions*, Neurocomputing. 413 (2020), pp. 210–216.
- [7] J.M. Chambers, C.L. Mallows, and B.W. Stuck, *A method for simulating stable random variables*, J. Am. Stat. Assoc. 71 (1976), pp. 340–344.
- [8] J.R. del Val, F. Simmross-Wattenberg, C.A. López, B. Rudis, B. Swihart, I. Rcpp, L. Rcpp, R.S.G. GSL, and M.B. Swihart, *Package ‘libstable4u’*, Commun. Stat.-Simul. Comput. 15 (2024), pp. 1109–1136.
- [9] J. Demongeot, K. Oshinubi, M. Rachdi, H. Seligmann, F. Thuderoz, and J. Waku, *Estimation of daily reproduction numbers during the covid-19 outbreak*, Computation. 9 (2021), pp. 109.
- [10] O. Diekmann, J.A.P. Heesterbeek, and J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol. 28 (1990), pp. 365–382.
- [11] W.H. DuMouchel, *Stable distributions in statistical inference: 2. Information from stably distributed samples*, J. Am. Stat. Assoc. 70 (1975), pp. 386–393.
- [12] E.F. Fama and R. Roll, *Parameter estimates for symmetric stable distributions*, J. Am. Stat. Assoc. 66 (1971), pp. 331–338.
- [13] J. Geweke, *Bayesian inference in econometric models using Monte Carlo integration*, Econometrica. 57 (1989), pp. 1317–1339.
- [14] X. He, E.H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y.C. Lau, J.Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B.J. Cowling, F. Li, and G.M. Leung, *Temporal dynamics in viral shedding and transmissibility of covid-19*, Nat. Med. 26 (2020), pp. 672–675.
- [15] M. Kanter, *Stable densities under change of scale and total variation inequalities*, Ann. Probab. 3 (1975), pp. 697–707.
- [16] S.M. Kogon and D.B. Williams, *Characteristic function based estimation of stable distribution parameters*, in: *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, 1998, pp. 311–338.
- [17] L. Komsta and F. Novomestky, *Package ‘moments’*, 2015; computer software available at <http://www.r-project.org>.
- [18] M. Kratz and S.I. Resnick, *The QQ-estimator and heavy tails*, Commun. Stat. Stoch. Models. 12 (1996), pp. 699–724.
- [19] E.E. Kuruoglu, *Density parameter estimation of skewed α -stable distributions*, IEEE. Trans. Signal. Process. 49 (2001), pp. 2192–2201.
- [20] M.J. Lombardi, *Bayesian inference for α -stable distributions: A random walk MCMC approach*, Comput. Stat. Data. Anal. 51 (2007), pp. 2688–2700.
- [21] S.M. Manou-Abi, Y. Slaoui, and J. Balicchi, *Estimation of some epidemiological parameters with the covid-19 data of mayotte*, Front. Appl. Math. Stat. 8 (2022), pp. 870080.
- [22] M. Matsui, *Fisher information matrix of general stable distributions close to the normal distribution*, preprint (2005). Available at arXiv math/0502559.
- [23] J.H. McCulloch, *Simple consistent estimators of stable distribution parameters*, Commun. Stat.-Simul. Comput. 15 (1986), pp. 1109–1136.
- [24] G.J. McLachlan, S.X. Lee, and S.I. Rathnayake, *Finite mixture models*, Annu. Rev. Stat. Appl. 6 (2019), pp. 355–378.
- [25] A. Nagaev and S. Shkol’nik, *Some properties of symmetric stable distributions close to the normal distribution*, Theory Probab. Appl. 33 (1989), pp. 139–144.
- [26] H. Nagatsuka, T. Kamakura, and N. Balakrishnan, *A consistent method of estimation for the three-parameter Weibull distribution*, Comput. Stat. Data. Anal. 58 (2013), pp. 210–226.
- [27] J.P. Nolan, *Parameterizations and modes of stable distributions*, Stat. Probab. Lett. 38 (1998), pp. 187–195.
- [28] J.P. Nolan, *Modeling with stable distributions*, in: *Univariate Stable Distributions: Models for Heavy Tailed Data*, 2020, pp. 25–52.

- [29] M. Nurminen and P. Mutanen, *Exact Bayesian analysis of two proportions*, Scand. J. Stat. 14 (1987), pp. 67–77.
- [30] T. Obadia, R. Haneef, and P.Y. Boëlle, *The R0 package: A toolbox to estimate reproduction numbers for epidemic outbreaks*, BMC Med. Inf. Decis. Making. 12 (2012), pp. 1–9.
- [31] K. Paczek, D. Jelito, M. Pitera, and A. Wyłomańska, *Estimation of stability index for symmetric α -stable distribution using quantile conditional variance ratios*, TEST 33 (2024), pp. 297–334.
- [32] M. Pitera, A. Chechkin, and A. Wyłomańska, *Goodness-of-fit test for α -stable distribution based on the quantile conditional variance statistics*, Stat. Methods Appt. 31 (2022), pp. 387–424.
- [33] S. Richardson and P.J. Green, *On Bayesian analysis of mixtures with an unknown number of components (with discussion)*, J. R. Stat. Soc. Ser. B: Stat. Methodol. 59 (1997), pp. 731–792.
- [34] G. Robinson, *Package 'fmstable'*, 2022; software available at <https://cran.r-project.org/web/packages/FMStable/FMStable.pdf>.
- [35] D. Salas-Gonzalez, E.E. Kuruoglu, and D.P. Ruiz, *Finite mixture of α -stable distributions*, Digit. Signal. Process. 19 (2009), pp. 250–264.
- [36] D. Salas-Gonzalez, E.E. Kuruoglu, and D.P. Ruiz, *Modelling with mixture of symmetric stable distributions using Gibbs sampling*, Signal. Processing. 90 (2010), pp. 774–783.
- [37] G. Samorodnitsky, M.S. Taqqu, and R. Linde, *Stable non-Gaussian random processes: Stochastic models with infinite variance*, Bull. Lond. Math. Soc. 28 (1996), pp. 554–556.
- [38] S.J. Sheather and M.C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation*, J. R. Stat. Soc.: Ser. B (Methodol.) 53 (1991), pp. 683–690.
- [39] K.P. Sinaga and M.S. Yang, *Unsupervised k-means clustering algorithm*, IEEE Access. 8 (2020), pp. 80716–80727.
- [40] Y. Slaoui, *Bandwidth selection for recursive kernel density estimators defined by stochastic approximation method*, J. Probab. Stat. 2014 (2014), ID 739640.
- [41] Y. Slaoui, *The stochastic approximation method for the estimation of a distribution function*, Math. Methods Stat. 23 (2014), pp. 306–325.
- [42] Å. Svensson, *A note on generation times in epidemic models*, Math. Biosci. 208 (2007), pp. 300–311.
- [43] M. Teimouri, *Statistical inference for stable distribution using EM algorithm*, preprint (2018). Available at arXiv:1811.04565, .
- [44] M. Teimouri, *Finite mixture of skewed sub-Gaussian stable distributions*, preprint (2022). Available at arXiv:2205.14067.
- [45] M. Teimouri, A. Mohammadpour, S. Nadarajah, M.M. Teimouri, S. Matrix, F. fBasics, and R. RUnit, *Package 'alphastable'*, 2022.
- [46] M. Teimouri, S. Rezakhah, and A. Mohammadpour, *EM algorithm for symmetric stable mixture model*, Commun. Stat.-Simul. Comput. 47 (2018), pp. 582–604.
- [47] G. Thakur and J. Saini, *Comparative study of iterative methods for solving non-linear equations*, J. Univ. Shanghai Sci. Technol. 23 (2021), pp. 858–866.
- [48] J. Waku, K. Oshinubi, and J. Demongeot, *Maximal reproduction number estimation and identification of transmission rate from the first inflection point of new infectious cases waves: Covid-19 outbreak example*, Math. Comput. Simul. 198 (2022), pp. 47–64.
- [49] J. Wallinga and M. Lipsitch, *How generation intervals shape the relationship between growth rates and reproductive numbers*, Proc. R. Soc. B: Biol. Sci. 274 (2007), pp. 599–604.
- [50] D. Wuertz, M. Maechler, and M.M. Maechler, *Package 'stabledist'*, 2016; software available at: <https://cran.r-project.org/web/packages/stabledist/stabledist.pdf>.
- [51] V.M. Zolotarev, *One Dimensional Stable Distributions*, American Mathematical Society, R. I. Providence, 1986.

Appendix. Additional tables and figures

We illustrate the data sets involved and perform a goodness-of-fit test that determines whether or not the sample data has skewness and kurtosis, to highlight that the distributions of our data sets are skewed, not symmetric, and have heavy tails.

Table A1. Goodness-of-fit test (p -value) of the Shapiro Wilk, Jarque–Bera, Kolmogorov–Smirnov and Anderson–Darling normality tests.

Dataset	Shapiro Wilk	Jarque–Bera	Kolmogorov–Smirnov	Anderson–Darling
SI data	0.0005387	0.006	$< 2.2e^{-16}$	$9.297e^{-5}$
N-acetyltransferase data	$< 2.2e^{-16}$	$2.776e^{-14}$	$< 2.776e^{-16}$	$< 2.776e^{-16}$

Table A2. Skewness and Kurtosis measures.

Dataset	Skewness	Kurtosis
SI dataset	0.8401043	3.57987
N-acetyltransferase data	1.197951	3.612564

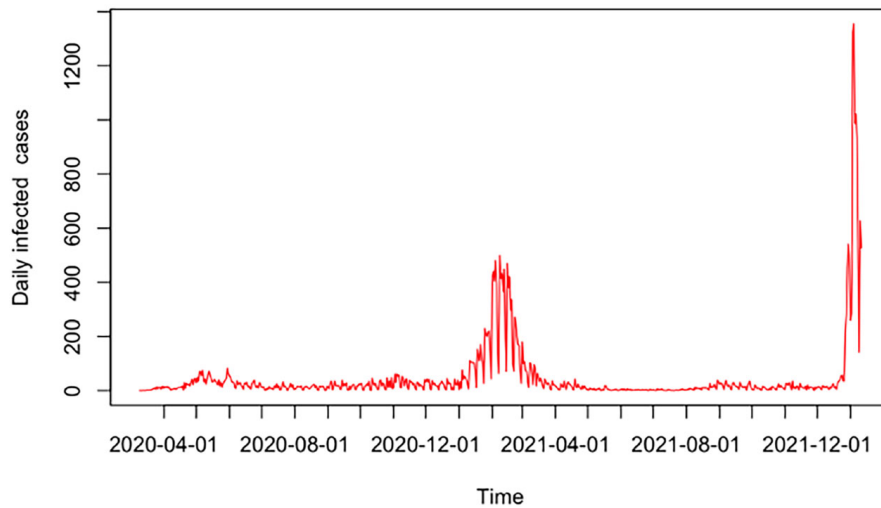


Figure A1. Daily reported cases from 13 March 2020 to 11 January 2022.

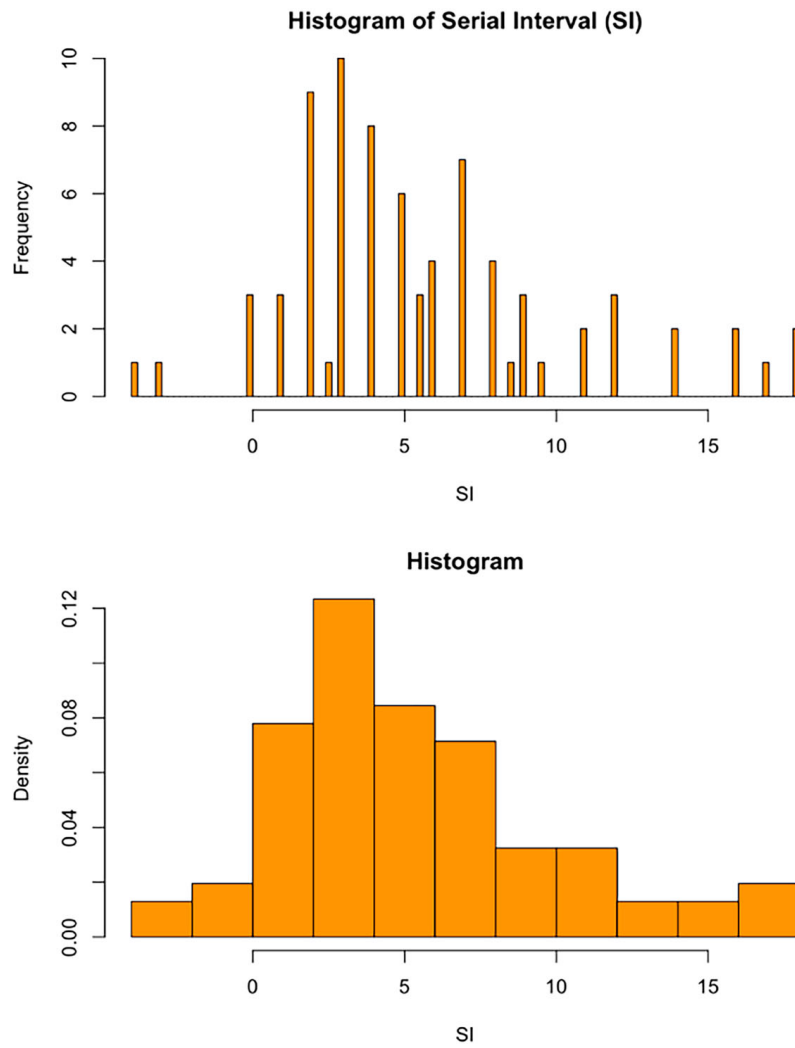


Figure A2. Serial interval data distribution from 77 infector-infectee transmission pairs [14].

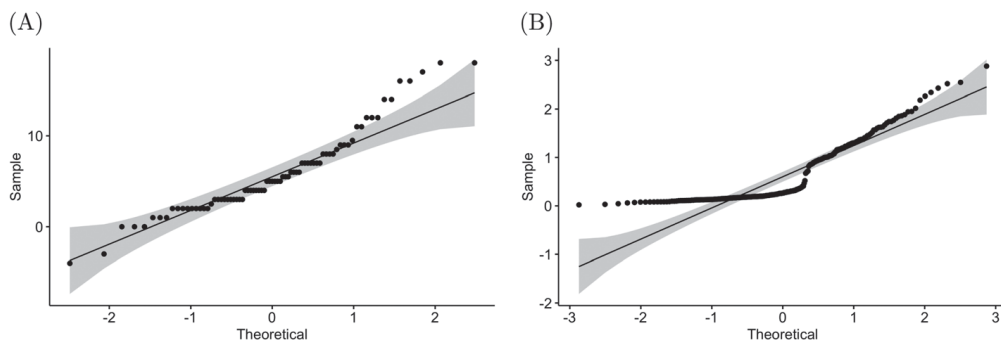


Figure A3. QQ-Plot of the SI data set distribution (A) and the distribution of the N-acetyltransferase activity data set (B).



Automatic geomorphological mapping using ground truth data with coverage sampling and random forest algorithms

Paul Aimé Latsouck Faye¹ · Elodie Brunel¹ · Thomas Claverie^{2,3} · Solym Mawaki Manou-Abi^{1,3,4} · Sophie Dabo-Niang⁵

Received: 1 February 2024 / Accepted: 26 May 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Marine geomorphological maps are useful to understand seafloor structure for example in the context of ecological studies, resources management or conservation planning. Although techniques to build such maps are increasingly sophisticated, manual techniques are still largely used. Automated approaches are needed to get reproducible maps in a reasonable time. This work provides statistical learning approaches based framework to build automatically geomorphological maps. We used bathymetric data to build Digital Bathymetric Model (DBM) and compute terrain attributes characteristic of seafloor geomorphology. Then, we used clustering based algorithms to select automatically ground truth locations from a reference geomorphological map manually made. Finally a supervised classification model random forest based was used to build predictive models for seafloor geomorphology typologies. Subsequently we studied the effect of DBM resolution, sample size and sampling method of the ground truth locations, in the quality of map production via a series of simulations. Results showed that the proposed framework allowed to build efficiently relevant seafloor geomorphological maps. The best compromise between the sampling effort and the quality of the resulting maps was obtained with 100 m DBM resolution, 200 data points sample size and using a complexity-dependent sampling method and led to a map matching at 90% the reference one.

Keywords Geomorphological map · Spatial modeling · Random forest classification · Digital bathymetric model · Terrain attributes · Lidar data

Introduction

Geomorphological maps are georeferenced delineation of morphological structure and surface composition of a studied land and/or seafloor (Otto and Smith 2013; Dramis et al. 2011; Pavlopoulos et al. 2009). Marine geomorphological maps are particularly crucial for resource management, conservation efforts, hazard assessment, protected area management, effective marine research campaign planning and various marine-related industrial works (Kienholz 1978; Bishop et al. 2012; Fukunaga et al. 2019; Browne et al. 2010). Such maps provide valuable information on the composition and structure of the seabed which, among other usage, is needed to generate habitat maps through the

identification and delineation of different benthic ecosystems like coral reefs, seagrass beds or various deep-sea communities (Pandian et al. 2009; Dramis et al. 2011; Wabnitz et al. 2008).

To produce geomorphological maps, different approaches are used (Siart et al. 2009; Hugenholtz et al. 2013). Widely used imagery techniques involve studying the patterns, textures, shapes, and color variations present in the imagery. This can be done manually by digitizing or tracing the features on the imagery or through automated or semi-automated image segmentation and classification techniques in Geographical Information System (GIS) using available image analysis tools. While imagery can be a valuable tool, it has a limited use for mapping deeply submerged geomorphological features due to water opacity. In such conditions, only acoustic approaches can provide usable data which are generally completed by punctual carefully located ground truthing observations using for example scuba-divers or submarine-divers observation, Remote Operated Vehicle (ROV) or Automatic Underwater Vehicle (AUV) picture

Communicated by: H. Babaie

Brunel Elodie, Claverie Thomas, Manou-Abi Solym and Dabo-Niang Sophie contributed equally to this work.

Extended author information available on the last page of the article

or videos, drop cameras or seabed sampling (Wynn et al. 2014; Locker et al. 2010). Subsequent treatments, required to generate maps with such data, will be to first propose a geomorphologic category for each ground truthing point (ie. Typology), then delineate surfaces of homogeneous typologies. Depending on whether ground truthing points are defined as categories or as quantitative data, interpolation methodologies within the surface to be mapped might take different forms.

Many semi-automated or automated approaches have been also proposed in recent decades to achieve objective, automated and repeatable approach to extract meaningful information using vast quantities of data (Summers et al. 2021). Object-Based Image Analysis (OBIA) which use bathymetric derivatives or a combination of bathymetric derivatives and backscatter to automatically segment the seafloor (Masetti et al. 2018; Argyropoulou et al. 2016; Lacharite et al. 2018; Koop et al. 2021; Dekavalla and Argialas 2017) are widely used. Bathymorphon / geomorphon-based classification (Jasiewicz and Stepinski 2013; Sowers et al. 2020; Ahn et al. 2023; Novaczek et al. 2019) and fuzzy logic scheme (Schmidt and Hewitt 2004; Lucieer and Lucieer 2009; Janowski et al. 2021) have been also investigated. The past 10 years, machine learning (Maschmeyer et al. 2019; Misiuk et al. 2021; Janowski et al. 2022; Sklar et al. 2024) and deep learning models (Behrens et al. 2018; Azarafza et al. 2023; Arhant et al. 2023) has increasingly been used for geomorphological mapping. Furtherthemore, the performance of these statistical learning methods has also been investigated in comparison with manually ones (Van der Meij et al. 2022; Diesing et al. 2014). In recent years, although these methods have proven their effectiveness, they are often combined with underwater imagery which superseded expert manual interpretation and are particularly costly for large scale mapping (Van der Meij et al. 2022; Cui et al. 2021; Galvez et al. 2022; Misiuk and Brown 2023; Breyer et al. 2023). As alternative, this work provides a clustering based algorithm for an optimal ground truth sampling and a learning-based approach for automatic geomorphological mapping. It focuses on the classical situation where experts use morphological map to define surfaces and ground truth measure to define typologies. But to that respect, further considerations need to be addressed: (i) the quality of bathymetric data required, (ii) the typology definition, and (iii) the methodology to use for creating geomorphological maps with such data.

(i) Bathymetry is recognized as essential when mapping geomorphology (Wilson et al. 2007; Lecours et al. 2016; Fukunaga et al. 2019). It is particularly useful for identifying and delineating submerged landforms that are not visible in aerial images. Modern bathymetric systems can

provide high-resolution data, capturing fine-scale details of the seafloor morphology. Their accuracy is often higher compared to aerial images, as it directly measures the water depth and seafloor elevation. However, the selection of the appropriate technology is crucial to ensure the acquisition of accurate and reliable data suited for an optimal geomorphological map production. More precisely equipment and material setup will depend on factors such as the scale and coverage needed, water depth, desired resolution, sampling effort and budgetary considerations. Compromise between the sampling effort and the quality of the data have to be made but little tools are available to choose the most optimal setup. Indeed very high-resolution data may be necessary for detailed local studies, while coarser resolution data might suffice for larger-scale regional or global analyses. In the present research, some examples to help on such decisions are proposed.

(ii) A typology is a description of one geomorphological category of seabed (in our case). However, to make a geomorphological map usable, comparable and understandable by a large international community, categories definition of typologies need to form a consensus. Our study is based on the Millennium Coral Reef Mapping Project (MCRMP) typologies. Initiated by the Institute for Marine Remote Sensing - University of South Florida (IMaRS/USF) in 2001 and continued since 2003 by reasearchers of the Institut de Recherche pour le Développement (IRD), the MCRMP has proposed a multi-level hierarchical structure (Andrefouet et al. 2004). MCRMP typologies were widely used as they enable several sites around the world to be compared on a thematically rich, homogeneous basis (Andrefouet and Dirberg 2006). These typologies provide a description of coral reef geomorphology distinguishing reef units such as slopes, flats, passes, terraces, lagoons, channels, etc.

(iii) Several approaches can be chosen to generate geomorphological maps from previously described data. The most classical approach is to manually draw typologies envelopes using GIS softwares (Minar and Evans 2008; Otto et al. 2018). However, it is tedious and poorly replicable if temporal changes need to be monitored. Automated approaches are of several kind. This work focuses on statistical learning based approaches for their ability to efficiently process and analyze large volumes of spatial data, to learn complex relationships present in the data leading to improved accuracy in produced maps (Stepinski et al. 2007; Siqueira et al. 2022; Van der Meij et al. 2022). More precisely four steps are followed: 1) generate a Digital Bathymetric Model (DBM) from bathymetric data, 2) compute terrain attributes from DBM on the entire surface studied, 3) train random forest based classification algorithm to match terrain attributes with ground

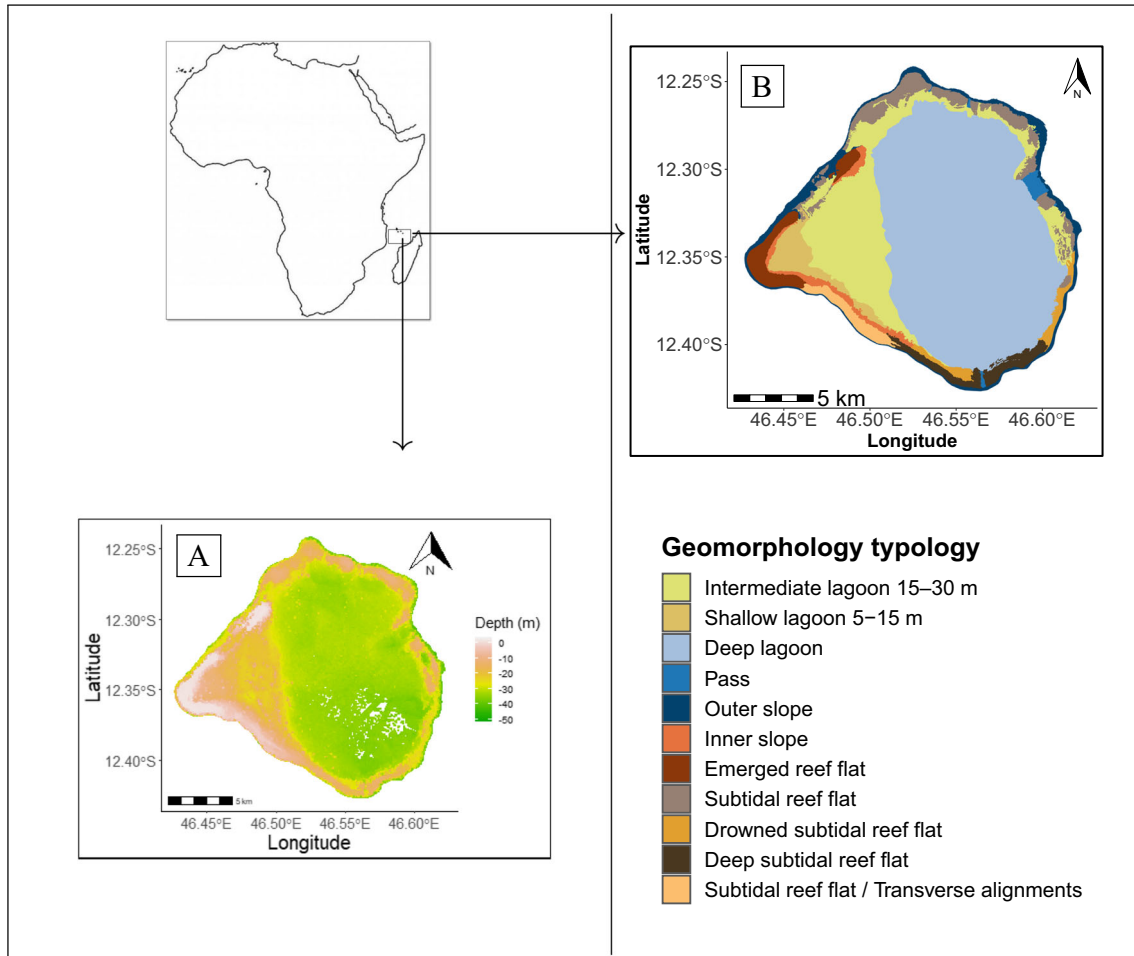


Fig. 1 Geyser is an atoll in the Western Indian ocean, specifically located in the northern Mozambique channel between Mayotte and the Gloriosos Islands and covering approximately 268 km². **A** Bathymetry data (in meters) collected by LIDAR technology and provided by

Hydrography and Oceanography Service of the Navy (SHOM). **B** Geomorphologic structure identified for the Geyser atoll, as a part of the EPICURE project (Roos et al. 2017)

truthing typologies and 4) predict typologies from the entire surface studied using the classification model generated.

The chosen methodology to generate geomorphological map was tested on a classical tropical reef feature: an atoll mapping from South Western Indian Ocean. Our approach is to propose a sensitivity analysis based on a comparison with an existing expert map while degrading DBM definition as proxy of data acquisition setup and varying the number of ground truth points as well as the methodology to select their locations as proxy of field effort. The ultimate goal is to supply tools to plan mapping field campaign using coverage sampling algorithms, codes to semi-automate geomorphological mapping procedure using random forest algorithm and propose metrics to evaluate the quality of the map generated across different resolutions. Recommendations to choose the DBM resolution, ground truth size and sites selection are also provided.

Materials and methods

Data

For this work, bathymetric data (depth measurements) collected between 2009 and 2010 on Geyser atoll with a surface area of approximately 268 km² are used (Fig. 1A). A set of 48.10⁶ data points were collected by LIDAR 1 m resolution calibrated. The depth range captured by this tool often reaches down to -30 m but due to exceptional very good water clarity conditions, bathymetric records on Geyser range between -50 and 4 meters (see their distribution in Fig. S1). This data is available on the Hydrography and Oceanography Service of the Navy website.¹

¹ <https://data.shom.fr/donnees>

An expert geomorphological map of the Geyser atoll available here² is also used. This scale-free map was produced by manual contouring, resulting from expert interpretation of hyperspectral images. 11 geomorphological typologies have been identified on Geyser (Roos et al. 2017).

General methodology

Here an automated scheme using bathymetry and some field typologies verifications also called ground truth to generate reproducible geomorphological maps is used (see Fig. S2 in the supplemental materials). The methodology is the following: 1) Using bathymetric data, a Digital Bathymetric Model (DBM) is created at a given resolution. 2) From the DBM, terrain attributes corresponding to raster layers which have the potential to influence seafloor geomorphology are computed. 3) Using a given coordinate sampling method, a given number of ground truth locations is drawn. By overlaying these locations with the expert map, geomorphological typologies are attributed to each ground truth points. 4) Using the training set of coupled terrain attributes and ground truth typologies, a recursive feature elimination algorithm based on a random forest classifier is used to select most relevant covariates. The latter are then used to train a random forest based predictive model for geomorphological typologies. This model is finally used to predict geomorphological typologies on the whole study site. 5) The quality of the map production was then evaluated using two types of performance criteria: the model performance (Balanced accuracy) and the matching between the generated map and the expert map (Match and Balanced match).

The robustness of our approach is also evaluated using a sensitivity analysis through simulation by varying bathymetric data definition, numbers of ground truth points and methodology to select their locations. This involved in examining five DBM resolutions (5 m, 25 m, 50 m, 100 m and 500 m), six sample sizes (50, 100, 200, 500, 700, 1000) and three sampling methods (two spatial coverage sampling methods denoted SCS-KMEANS and SCS-CLARA, and a complexity-dependent sampling method denoted CDS). For each of these combinations ($5 \times 6 \times 3$), 30 replicates of geomorphological maps are generated by replicating the steps 3-5 described above.

Digital bathymetric model creation

The raw bathymetric dataset contains approximately 48.10^6 data points giving latitude, longitude and depth values. To avoid numerical issues due to redundant observations, this dataset was spatially sub-sampled and data points are kept as

² <https://sextant.ifremer.fr/geonetwork/srv/api/records/21232c12-e409-4136-a24a-78c346518cfa>

homogeneous as possible using the *buffer.points* function in the supplemental materials (Roberts 2015). The sub-sampled data are such that each retained data point is at least at a distance of 5 m from one another. Following this procedure, the new dataset of approximately 8.10^6 data points obtained is used to generate the DBM which consists in the creation of a square grid of T^2 cells or rasters. The cell size define the DBM resolution. Different resolutions (5 m, 25 m, 50 m, 100 m, 500 m)³ are created using the *dbm* function proposed in the supplemental materials. The bathymetry or depth value of each cell is calculated by taking the average of all the depths inside the same cell.

Furthermore, some deep areas are poorly sampled, due to the lack of signal return on the LIDAR sensor (see white zone in the middle of Fig. 1A). To get a depth measure for each ground truth location selected using one of the automatic sampling methodology presented further, depth measure on each raster of the study area were required. Thus for each DBM resolution, missing depth of empty cells in this zone are interpolated using an ordinary kriging model via the *ok.dbm* function provided in the supplemental materials. The kriging method is not described in details since the problem of missing data is out of the scope of the paper (Cressie 1988). In contrast, empirical results provided in the supplemental materials show that it performs better than five others spatial interpolation methods for our data Table S1. This imputation method allows to complete bathymetric data on these cells.

Terrain attributes calculation

It consists in quantifying predictors for seafloor geomorphology. For each DBM resolution, terrain attributes were calculated using a moving routine from the DBM. More precisely, let us denote the depth $D_{i,j}$ of a given cell with $(i, j) \in \llbracket 1, T \rrbracket \times \llbracket 1, T \rrbracket$ and consider the depth $D_{i+k,j+l}$ of the neighboring cells with $(k, l) \in \mathcal{D}_3 = \llbracket -1, 1 \rrbracket \times \llbracket -1, 1 \rrbracket$. These cells are defining the 3×3 window on which the terrain attributes are defined. Note that it is possible to generalize the study to square window with size greater than 3. Selected terrain attributes can be organised into three groups: Slope (*Slope*) and orientation (*Aspect*) measures, terrain variability measures such as Roughness (*Roughness*), Terrain ruggedness index (*TRI*) and Vector ruggedness measure (*VRM*) and curvature and relative position measures such as Profile convexity (*prof_c*), Planform convexity (*planc*) and Bathymetric position index (*BPI*).

³ The 5 m resolution is the one used by biologists during field verification campaigns. We then looked for a very high resolution at which the geomorphological maps produced were degraded because they were too pixelated. 500 m seemed to be a good choice. Between these two resolutions, we empirically searched for intermediate resolutions enabling us to obtain “significantly” different maps. Hence the 25 m, 50 m and 100 m resolutions were retained.

Slope has been widely recognized as an important factor for determining benthic habitat and colonization and has been used in many marine studies (Copeland et al. 2013; Fukunaga et al. 2019; Sterne et al. 2020). Like the magnitude of the steepest drop in depth, *Slope* (in degrees) is derived from rates of change in x (longitude) and y (latitude) directions and is calculated as follows, for each cell $(i, j) \in \llbracket 1, T \rrbracket \times \llbracket 1, T \rrbracket$,

$$Slope_{i,j} = \frac{180}{\pi} \arctan \left(\sqrt{\left(\frac{\partial D_{i,j}}{\partial x}\right)^2 + \left(\frac{\partial D_{i,j}}{\partial y}\right)^2} \right). \quad (1)$$

where

$$\begin{cases} \frac{\partial D_{i,j}}{\partial x} = [(D_{i+1,j+1} + 2D_{i+1,j} + D_{i+1,j-1}) - (D_{i-1,j+1} + 2D_{i-1,j} + D_{i-1,j-1})] / 8\Delta. \\ \frac{\partial D_{i,j}}{\partial y} = [(D_{i+1,j+1} + 2D_{i,j+1} + D_{i-1,j+1}) - (D_{i+1,j-1} + 2D_{i,j-1} + D_{i-1,j-1})] / 8\Delta. \end{cases}$$

where the real number Δ stands for the cell size of the grid.

The orientation measure (*Aspect*) gives the exposure of a given area to such water waves and is often used in the calculation of others parameters that directly influence habitat (Wilson et al. 2007). *Aspect* (in degrees) is the compass direction of the steepest drop in depth and is calculated as follows, for each cell $(i, j) \in \llbracket 1, T \rrbracket \times \llbracket 1, T \rrbracket$,

$$Aspect_{i,j} = 180 + \frac{180}{\pi} \arctan \left(\frac{\partial D_{i,j}}{\partial x} + \frac{\partial D_{i,j}}{\partial y} \right). \quad (2)$$

The roughness measure (*Roughness*) is a critical factor affecting ecological and physical processes on the reef (Leon et al. 2015; Dartnell 2000). It corresponds to the difference between the maximum and minimum depth values over a 3×3 window and is defined for each cell $(i, j) \in \llbracket 1, T \rrbracket \times \llbracket 1, T \rrbracket$ as follows:

$$Roughness_{i,j} = \max_{k,l \in \mathcal{D}_3} (D_{i+k,j+l}) - \min_{k,l \in \mathcal{D}_3} (D_{i+k,j+l}). \quad (3)$$

The Terrain Ruggedness Index (*TRI*), is a terrestrial ruggedness measure (Riley et al. 1999) that was adapted to bathymetry data to highlight morphological heterogeneity (Valentine et al. 2004; Rozycka et al. 2017). It is defined as the mean of the absolute differences between the depth value of a cell and the one of its neighboring cells, for each cell $(i, j) \in \llbracket 1, T \rrbracket \times \llbracket 1, T \rrbracket$ as follows:

$$TRI_{i,j} = \frac{\sum_{k,l \in \mathcal{D}_3} |D_{i+k,j+l} - D_{i,j}|}{(3^2 - 1)}. \quad (4)$$

The Vector Ruggedness Measure (*VRM*) quantifies terrain ruggedness : slope and aspect are decomposed into 3-dimensional vector components using standard vector analysis in a user-specified moving 3×3 window. The vector ruggedness measure is dimensionless because it involves sine and cosine of the slope and aspect measures and its values range from 0 to 1 corresponding to flat regions to rugged ones. Its mathematical definition is omitted to avoid technicalities and details can be found in Sappington et al. (2007).

The Curvature position may also be linked to the nature of the seabed. It helps to delimit regions of distinct habitat by identifying boundaries in the character of the terrain (Wilson et al. 2007). Bathymetric Position Index (*BPI*), the marine version of the topographic position index, quantifies where a location on a bathymetric surface is relative to the overall seascape (Mata et al. 2021). It provides an indication of whether any particular pixel forms part of a positive (e.g., crest) or negative (e.g., trough) feature of the surrounding terrain (Lundblad et al. 2006; Wilson et al. 2007). It is calculated using the following formula, for each cell $(i, j) \in \llbracket 1, T \rrbracket \times \llbracket 1, T \rrbracket$:

$$BPI_{i,j} = D_{i,j} - \frac{\sum_{k,l \in \mathcal{D}_3} |D_{i+k,j+l} - D_{i,j}|}{(3^2 - 1)}. \quad (5)$$

According to Evans (1980), Profile convexity (*Prof_c*) is the rate of change of *Slope* and Plan convexity (*Plan_c*) is the rate of change of *Aspect*. Negative values in the *Prof_c* indicate the surface is upwardly convex whereas, positive values indicate that the surface is upwardly concave. Positive values in the *Plan_c* means the surface is laterally convex and negative values indicate that the surface is laterally concave. Several methods exist for numerical approximations of these metrics, based on a quadratic form representation f of the surface (Florinsky 1998; Horn 1981; Evans 1980; Zevenbergen and Thorne 1987). Numerical implementation of Zevenbergen and Thorne (1987) method's is used in this study ; details can be found in Florinsky (1998).

All the computed terrain attributes, depth and geographic coordinates (longitude, latitude) are then stacked to form a multilayer grid of predictors called features or covariates in the sequel Table 1.

Sampling

To build predictive models for geomorphological typologies using a statistical learning approach, a training dataset is required. This one must contain a finite number of ground truth locations and a set of predictive covariates for geomorphological typologies at these locations. In this section, three alternative clustering based sampling methods are proposed to draw automatically these locations inside a given study

Table 1 Terrain attributes computed from the DBM and functions used to do such calculations in R software

Terrain attributes	Reference	Function / R Package
Slope and Aspect		
Slope	(Horn 1981)	terrain / raster
Aspect	(Horn 1981)	terrain / raster
Terrain Variability		
Roughness	(Dartnell 2000)	terrain / raster
Terrain Ruggedness Index (<i>TRI</i>)	(Wilson et al. 2007)	terrain / raster
Vector Ruggedness Measure (<i>VRM</i>)	(Ilich et al. 2023)	VRM / MultiscaleDTM
Curvature and relative position		
Profile Curvature (<i>ProfC</i>)	(Zevenbergen and Thorne 1987)	Curvature / spatialEco
Planform Curvature (<i>PlanC</i>)	(Zevenbergen and Thorne 1987)	Curvature / spatialEco
Bathymetric Position Index (<i>BPI</i>)	(Ilich et al. 2023)	BPI / MultiscaleDTM

area. All these methods use cell's centers of a regular grid and choose among the covariates earlier mentioned depending on the chosen algorithm.

Spatial Coverage Sampling using k-means clustering algorithm (SCS-KMEANS)

The basic idea of Spatial Coverage Sampling (SCS) is to draw uniformly sampling locations over the study area. It has been shown that SCS on a study area can be achieved by k-means clustering algorithm (Hartigan 1975). This consists in grouping cell's centers of a regular grid on this area using their spatial coordinates as covariates. Note that this regular grid can be the DBM one as long as it does not lead to computational deadlock, otherwise it can be replaced by a raster grid with lower resolution. The final solution of this partition gives the sampling locations and is determined by minimizing a geometric criterion, the mean squared shortest distance between the clusters centroids and the grid cell's centers (Royle and Nychka 1998; Brus et al. 2006). For the implementation, a k-means algorithm for equal area partitioning is used (Brus 2019). The *scsKM* function provided in the supplemental materials can be used to achieve this.

Spatial Coverage Sampling using CLARA algorithm (SCS-CLARA)

K-means clustering approach is time and storage consuming, especially for high DBM resolution. In such case, CLARA algorithm approach could be an alternative. The CLARA algorithm is an extension of the Partitioning Around Medoids (PAM) methods (Kaufman and Rousseeuw 1975) to deal with data containing a large number of objects (more than several thousand observations) in order to reduce computing time and storage problem. Medoids (M_i)_{*i*=1}^{*k*} in a PAM, *k* being the desired number of clusters *C_i*, are cells that minimize their distance to other cell's centers of the cluster. The CLARA algorithm generates *j* ∈ \mathbf{N}^* random samples of size *n* (*n* < *T*²) on individuals, applies a PAM on these samples one after the other, then evaluates the partition quality on

each of them by calculating the average global dissimilarity on the complete dataset as follows:

$$\sum_{i=1}^k \sum_{x_c \in C_i} \frac{d(x_c, M_i)}{T^2} \quad (6)$$

If this dissimilarity is lower than the previous found one, it considers this solution and its *k* medoids as the best current solution. The *scsCLARA* function in the supplemental materials can be used to choose ground truth locations.

Complexity-Dependent Sampling using CLARA algorithm (CDS)

Terrain morphology can sometime be marked by accidented region where many typologies are observed closed to each other. This is particularly true with coral reef region where for example sand flat, reef slope, reef flat, etc typologies can be close to each other. In such situations, homogeneous sampling would lead to miss many typologies unless a lot of points are requested leading to other problems such as important unbalanced typologies distribution among locations (Brus 2019). To account for such typologies distribution, ground truth locations are agglomerated around zone of higher ground complexity (i.e. complex terrain). Thus the CDS method is introduced in order to take advantage of these covariates available for each DBM resolution. CDS aims to distribute sampling locations around most heterogeneous or complex areas. It uses DBM's cells as individuals and unlike the SCS which uses the spatial coordinates, CDS's covariates are chosen among terrain attributes and depth. Depth and Roughness measures are considered in this study. CDS cannot be computed using the k-means clustering algorithm for high DBM resolutions because DBM cells are mandatory to get covariates. This is why the CLARA algorithm is only used. The *cdCLARA* function in the supplemental materials can be used to this end (Fig. 2).

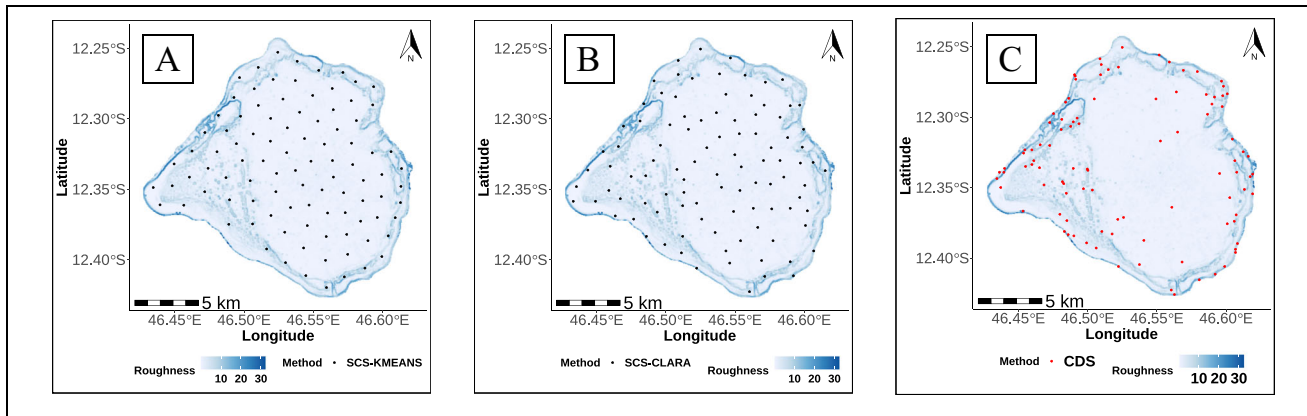


Fig. 2 Example of 100 ground truth locations drawn with the different methods: **A** Spatial Coverage Sampling using K-means clustering algorithm (SCS-KMEANS), **B** Spatial Coverage Sampling using CLARA

algorithm (SCS-CLARA) **C** and Complexity-dependant sampling using a CLARA algorithm (CDS). Roughness and Depth are used to guide the complexity-dependant sampling

Geomorphology mapping using Random Forest algorithm

To map geomorphological typologies over a whole study area, a supervised classification approach is considered. This work was divided in two steps: a first step to train algorithm and a second step to predict typologies based on trained algorithm. For the training part, each location is generated by one of the previously described sampling methods and typologies were attributed using the expert map Fig. 1B. These located typologies were the target variable, locations coordinates and the corresponding depth and terrain attributes were used as covariates and both formed the training dataset. In this process, a feature selection scheme random forest based is used to subset the most relevant covariates of the training set. Then a final model is fitted using the selected covariates. After this first step completed, the second step consisted in using the fitted model to predict the most suitable typology over the whole study area.

The generic principle of Random Forest

The tree based Random Forest (RF) algorithm can be used for a classification task (Breiman 2001; Biau and Scornet 2016). Using the Bootstrap AGGREGATING (bagging) principle, RF increases the diversity of the trees by making them grow from different randomly drawn (with replacement) training datasets from the original dataset (Breiman 1996). At each node of each tree, Rf selects a random subset of features and search for the best split for the node. To classify a new case once the forest is completed, the typology having the most votes over all the trees is retained.

Model training Using a cross-validation scheme with 3 repeats, the training samples are split into 3 folds. To train a RF model, three hyperparameters were tuned: the number of trees (*Ntrees*) of the forest, the number of features used at each node (*Mtry*) and the minimum number of data points

at the terminal node of each tree (*nodesize*). Indeed, the *Ntrees* hyperparameter is not tunable in the classical sense but should be set sufficiently high (Diaz-Urriarte and Alvarez de Andres 2006; Oshiro et al. 2012; Probst et al. 2019). The default value of 500 trees is used. The *nodesize* hyperparameter has been set to 1 for classification task because it generally provides good results (Diaz-Urriarte and Alvarez de Andres 2006). The *Mtry* hyperparameter was tuned among fifteen values of hyperparameters chosen automatically by the function *tuneLength*. A random search optimization strategy which defines a search space as a bounded domain of parameter values and randomly sample points in that domain were used to find the optimal *Mtry* considering the the resulting Accuracy (Grandini et al. 2020). Models are fitted by repeatedly leaving out one of the folds and performance are determined by predicting on the fold left out. The *train* function of the *caret* R package were used to this end.

Terrain attributes selection

By selecting the most relevant terrain attributes as covariates, the risk of over fitting can be reduced and the model's generalization ability improved. Variable or feature importance measures are usually used to rank or select variables. Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA) are two well-known random forest variable importance measures (Guyon and Elisseeff 2020; Breiman 2001; Biau and Scornet 2016). In this study, the MDA measure also called permutation importance in Breiman's original random forest is chosen since it seems to exhibit less bias than MDI in presence of correlated features (Strobl et al. 2008; Breiman 2001). Roughly speaking, the MDA measure consists in shuffling values of a given covariate *j* in the test data or out-of-bag data (that is data excluded from the bootstrap sample used to construct the tree) and then computes the difference between the error on the permuted test set and the

original test set. More precisely, for each tree t among the n_{tree} trees of the RF, MDA uses the out-of-bag data to compute a prediction error OOB_t . Then, permuting the values of the j^{th} feature in the out-of-bag data, a prediction error OOB_t^j is computed by using the permuted out-of-bag data. The permutation importance of the feature j is thus defined by $MDA_j = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} (OOB_t - OOB_t^j)$.

To achieve feature selection via the MDA measure, the backward Recursive Feature Elimination (RFE) algorithm (Guyon et al. 2002) implemented in the *rfe*⁴ function of the *caret* R package is used. This algorithm is based on assessing MDA's importance. MDA is computed by iteratively permuting the values of each input covariate and measuring the resulting drop in prediction accuracy. The feature with the minimum MDA value, representing the least important feature, is systematically removed from the input set. Subsequently, a new RF model is trained using this reduced set of features. This process is continued until the minimal set of input features that yielded optimal Accuracy (Grandini et al. 2020) is obtained. To improve the performance of feature selection with RFE, a repeated 3-fold cross-validation with 3 repeats is used. The *createFolds* function of the *caret* package allowed to split data into training and test sets. This function carries a random sampling within geomorphological typologies in order to balance the classes distributions within the split.

Once the model is trained, typologies can be predicted across the entire study area. Typologies predictions are made using the *predict.train* function of the *caret* R package (Kuhn 2019).

Performance criteria

To assess model's performance, a *Balanced Accuracy* (*BA*) metric were calculated using the confusion matrix obtained from each model (Grandini et al. 2020). It consists for each typology (class) k , to calculate a Recall score measuring the ability of a model to find all the positive units for this class as follows:

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \quad (7)$$

True Positives (TP) are observations predicted to belong to the reference class when they really do, and False Negatives (FN) are observations predicted to not belong to the reference class when they really do. *BA* gives an average measure of this concept using the arithmetic mean of Recall across all classes :

$$BA = \frac{\sum_{k=1}^K Recall_k}{K} \quad (8)$$

⁴ <https://topepo.github.io/caret/recursive-feature-elimination.html>

where K is the total number of class k .

The number of unsampled typologies is used as an indicator of the efficiency of the sampling method to visit all geomorphologic typologies present in the study area.

To assess the consistency of predictions, the map produced at the different DBM resolution are compared to Expert map. For this purpose, the expert map is first discretized into "pixels" size of 5 m resolution (ie. the smallest resolution used in this study). Then the predicted map grid is disaggregated to match the expert map resolution using *disagRast* function in the supplemental materials. To compare two maps, two metrics are calculated using the confusion matrix between the two: *Match* and *Balanced Match*.

The *Match* metric is a simple comparison between the two maps. The *Match* metric corresponding to the proportion of cells identically labelled is calculated as follows:

$$Match = \frac{N_+}{N_+ + N_-} \quad (9)$$

Where N_+ is the total number of cells where typologies match between the two maps, and N_- is the total number of cells that do not match. The *matchRast* function is proposed for the calculation of this metric. However such comparison might hide prediction problems on some small surface typologies, totally dominated by high surface coverage of some typologies. To take account of such fact, the *Balanced Match* is a sort of ponderated *Match* removing the surface dominant effect that some typologies might have on others. A second metric denoted *Balanced Match* (*BM*) inspired by the *BA* defined in Eq. 8 and using the confusion matrix between a predicted map and the expert map is also calculated. Note that after disaggregating a predicted map, some cells centers around the study site borders may be located outside. These border effects are handled by proportioning *BA* to the proportion of cells L labelled after the disaggregation of a predicted map. *BM* is calculated as follows:

$$BM = L \cdot \frac{\sum_{k=1}^K Recall_k}{K} \quad (10)$$

The *balmatchRast* function provided in the supplemental materials is used for the *BM* calculation.

Results

Effect of ground truth sampling methodologies on typologies sampled

Our first methodological assessment was to evaluate the number of missed typologies on the expert map depending on ground truth sampling techniques. Unsurprisingly, the more

data points sampled, the greater the chance of sampling all typologies present in the study site, regardless of the sampling method and the resolution of the DBM (Fig. 3). Whatever the sampling method and the DBM resolution, small sample size (50 and 100 data points) do not allow to sample all typologies. For these sample sizes, between 1 and 4 typologies are never sampled. It can also be noticed that for 500 data points sampled and more, all typologies are sampled no matter the sampling method and the resolution of the DBM.

Assessment of model performance based on input data

The *Balanced Accuracy* criteria helps assessing model performance. The larger the sample size, the more precise these measures (ie. smaller standard errors; Fig. 4). For small sample sizes (50 and 100 data points), sampling methods give comparable results or even slightly better results for SCS-KMEANS than others sampling methods. In contrast, from 200 points upwards, complexity-dependent sampling tends to give better results than SCS-CLARA and SCS-KMEANS which give comparable results. This result is confirmed as the sample size increases, with a widening gap among sampling

methodology means and decreasing standard errors (see Fig. S3 and Table S3 for further details).

Assessment of selected terrain attributes

A set of features is selected during the feature selection step of the modeling process and general outputs are summarized. Only results for 100, 200, 500 data points sample sizes generated at 100 m DBM resolution using SCS-CLARA method and CDS method are shown her because they are representative of all results (Fig. 5). Geographic coordinates (*Latitude* and *Longitude*) and *Depth* are almost always selected regardless the DBM resolution and the sampling method. Terrain variability attributes groups (*Roughness*, *VRM* and *TRI*) that are selected a little more than half of the time.

Evaluation of the quality of produced maps

The *Match* criteria evaluates the consistency of predictions according to the expert map. As previously seen with the *Balanced Accuracy* standard errors values, *Match* errors are non negligible for small sample sizes and decrease when

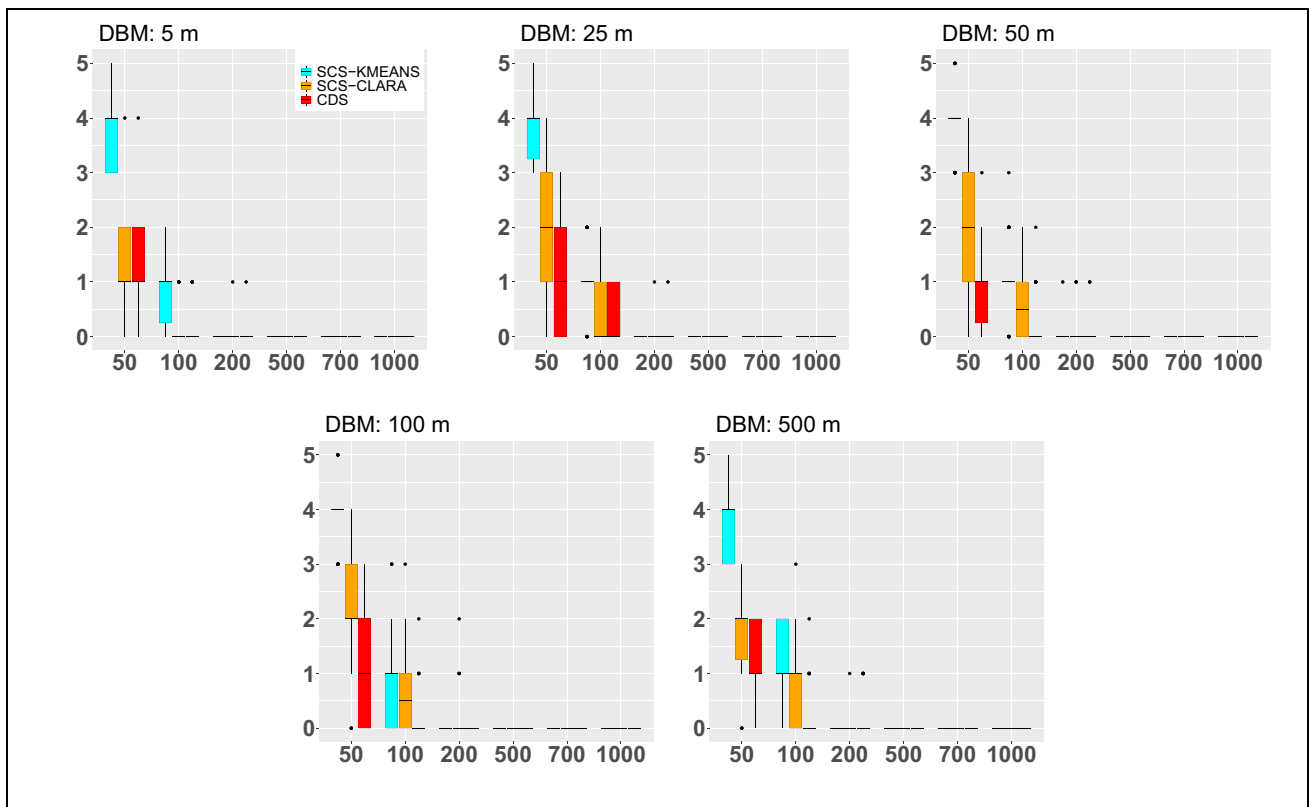


Fig. 3 Boxplot representing the *Number of missing typologies* metric (y-axis) for different sample size (x-axis), different sampling methods (CDS, SCS-CLARA and SCS-KMEANS) and different bathy-

metric model resolutions (5 m, 25 m, 50 m, 100 m, 500 m). Each sampling conditions were replicated 30 times and represented as default R boxplot settings

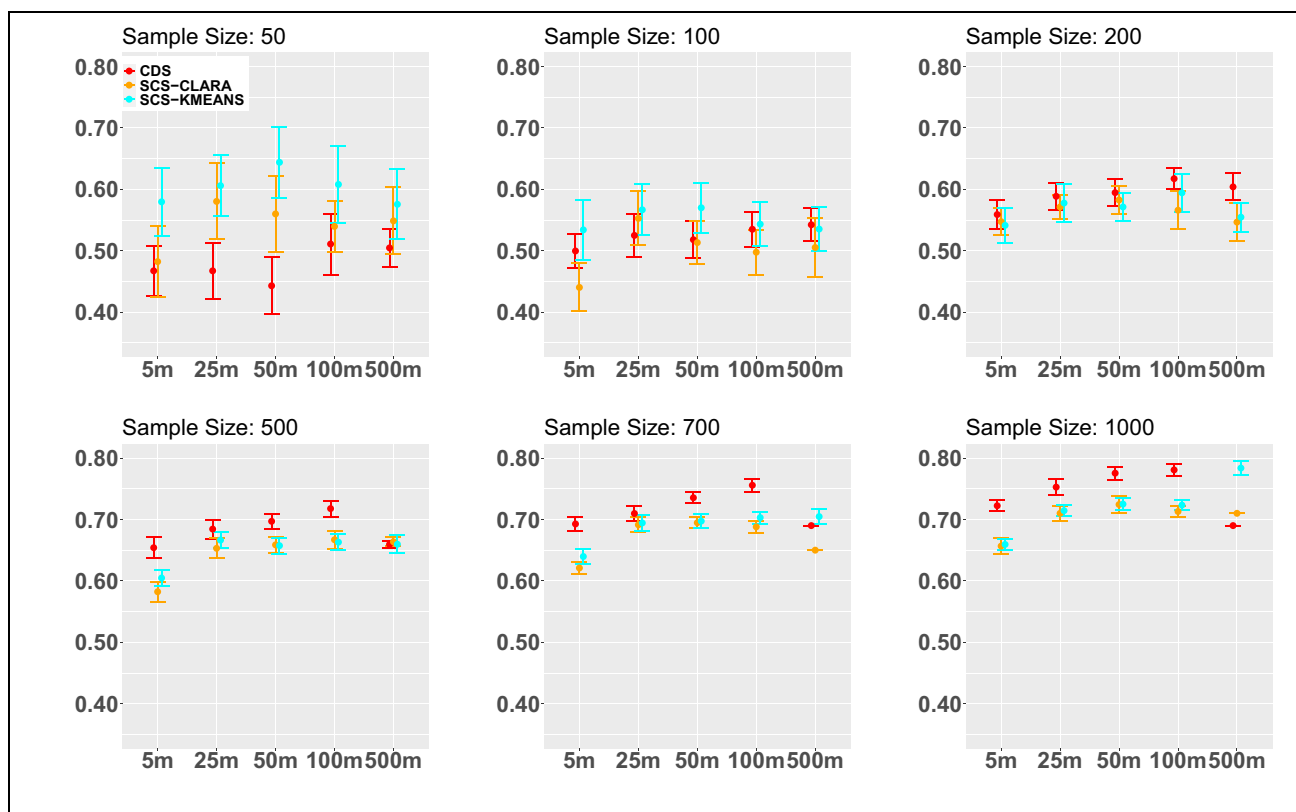


Fig. 4 Mean (+/- standard error) over 30 draws of the *Balanced Accuracy* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

DBM resolutions increase (Fig. 6; Fig. S3 and Table S4 for further details). In addition, lower *Match* values are recorded at 5 m and 500 m DBM resolutions and higher *Match* values were seen for intermediate DBM resolutions (25 m, 50 m, 100 m) regardless the sample size.

When the sample size and the DBM resolution increase, *Balanced Match* values become more accurate as shown by a decrease of standard errors (Fig. 7; Fig. S3 and Table S5 for further details). CDS method gives better results than others spatial coverage sampling methods for any sample size and any DBM resolution. SCS-CLARA and SCS-KMEANS give comparable results according this criteria. Comparing results among DBM resolutions, best results are achieved with 50 m and 100 m DBM resolutions.

In Fig. 8 two maps generated with data points sampled using CDS methodology are restituted: (A) 50 m DBM resolution using 1000 data points and (B) 100 m DBM resolution using 200 data points. All typologies are sampled and predicted in both cases. Results show that performance criteria measured are better for the (A') case than the (B') case (see (C)). Indeed, some small surface typologies like Drowned subtidal reef flat and inner slope are better predicted and predictions errors on typologies transition areas less important for (A) than for (B).

Discussion

This methodology enabled us to successfully reconstruct an expert geomorphological map. According to the *Number of missing typologies* metric, 200 data points are enough to sample almost all the typologies. These locations can be chosen using CDS which performs slightly better than spatial coverage sampling methods (SCS-KMEANS and SCS-CLARA). Results between the different DBM resolutions are comparable for small sample sizes considering the *Balanced Accuracy* metric. But, for 200 data points and more, 100 m DBM resolution gives clearly better results than others resolutions. The *Match* metric supports this finding where 50 m and 100 m DBM resolutions gives much better results than others DBM resolutions whatever the number of data points. Considering the *Balanced Match* criteria which was introduced to contrast the *Match* criteria taking into account typologies surface imbalances, 50 m DBM resolution gives a slight better performance than the 100 m DBM resolution, but considering the sampling effort, it would be preferable to use the 100 m DBM. Indeed, while the sampling effort is multiplied by 5 from (B) to (A), the performance recorded is just a little bit better. We have also seen that the precision of the maps produced is sensitive to the resolution

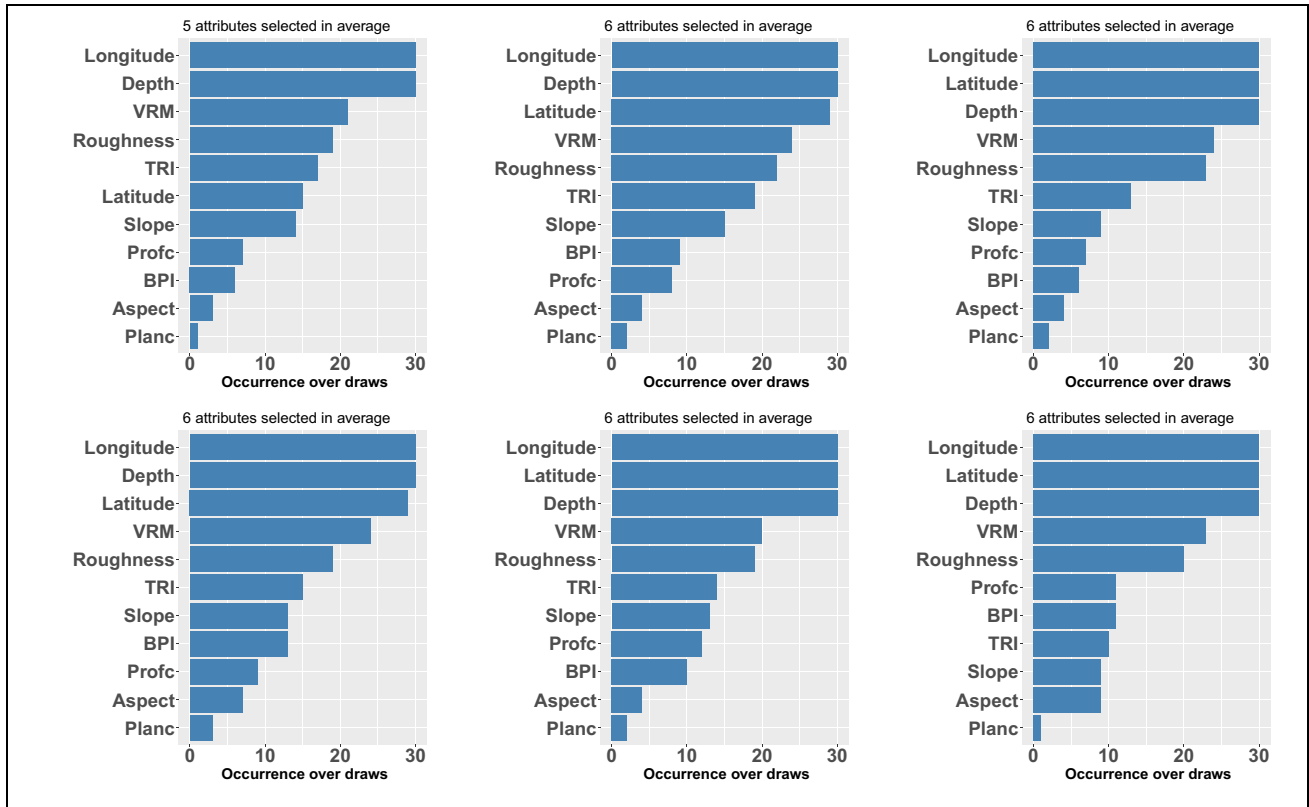


Fig. 5 Ordered occurrence over 30 draws of the selected terrain attributes by the RFE algorithm. Here an example with 100, 200, 500 data points sample sizes (from left to right) generated at 100 m DBM

resolution using SCS-CLARA method (top) and CDS method (bottom). The average number of attributes selected is noted above each graph (cf. Terrain attribute section for their definition)

of the DBM, the number and locations of the ground truth selection. Thus, reproducible methodologies with associated codes to evaluate their qualities are proposed. In this section, choices on data sampling to generate such maps are discussed. Then strong and weak points of the modeling approach and alternative to enhance such work are addressed.

Optimal data acquisition parameters

Creation of a submarine geomorphological map appears, among other consideration, to be constrained between quantity and quality of initial data and cost to acquire and process them. In the present work two kind of data fall in such compromise: the bathymetric data and the ground truth data points.

For the bathymetric data acquisition, high resolution data require advanced technologies, longer survey durations, important storage and sophisticated tools for manipulation which all contribute to higher costs. The results of this study show that a lower DBM resolution do not necessarily lead to the best geomorphological map. The 100 m DBM resolution considered in this study appears to strike a good

balance between detailed geomorphological maps obtained (Figs. 4, 6, and 7) with low DBM resolutions (5 and 25 m) and large scale geomorphological maps considering (500 m DBM resolution). Indeed, maps at 5 m, 25 m and 50 m DBM resolutions provide a reasonably detailed view of the seafloor and allow the identification of important features but contain a significant amount of noise or small-scale variability that may not be relevant in constructing a geomorphological map. On the other hand, maps of 500 m resolution provide very generalized view of the seafloor representing a significantly coarser scale focusing on major landforms and regional-scale geomorphological patterns but missing important features required to build such maps (eg. reef patches, pass or canyons ...). The 100 m resolution was also considered suitable in previous studies for various applications, such as regional planning, environmental assessments, and natural hazard evaluations (Curie et al 2007; Dong et al. 2019).

Ground truth data acquisition is often acquired by specialists through scuba diving or snorkling (shallow) or using submarine, ROV or drop camera systems allowing to view and characterise seafloor typologies on specific location. In that respect, the number of location is directly correlated to

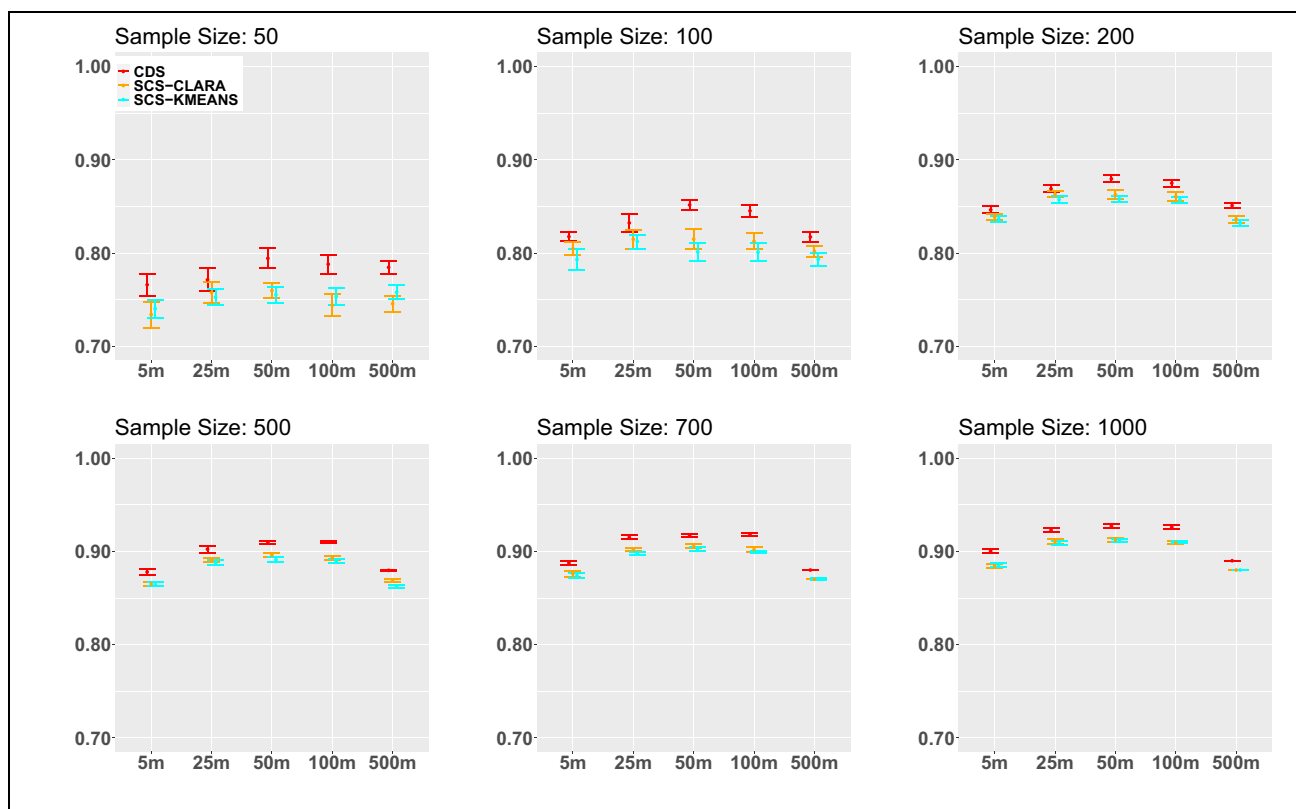


Fig. 6 Mean (\pm standard error) over 30 draws of the *Match* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

cost of data gathering and therefore need to be balanced. Furthermore, if the number of data point is directly linked to survey cost, the location of sampling on an heterogeneous seafloor structure could be related to data quality (ie. enhanced typology sampling diversity above oversampling of common typologies). To test such compromises, effects number of data points and methodologies to locate them (SCS vs CDS approach) on overall map quality production are studied. The choice between homogeneous distribution vs seafloor complexity dependent approach came from the fact that various typologies might agglomerate around location of complex structures (e.g. barrier reef, patch reef, ...). Figure 2 illustrates this point and show that CDS method amplify sampling in area of important typologies diversity. This explains why *Number of missing typologies* criteria results obtained with CDS are better than homogeneous sampling methods (Fig. 3). However, such approach has the inconvenient that a bathymetric dataset is required prior planning ground truth sampling campaign. Another point of attention is that attributes used as covariates for CDS algorithms should be carefully selected. This study according to the context of used data (coral reef habitats) focuses on metrics promoted in literature (Adey 1966; Adey and Macintyre 1973; Battistini 1975; Minnery et al. 1985) and on empirical

results from this work in the supplemental materials with a supervised classification of geomorphologic typologies based on the terrain attributes presented in Table 1. Seafloor depth and roughness were retained as the most relevant.

Evaluating effect of bathymetric data points density and number and location of ground truth points was done by comparing produced map to an existing expert manually made map. The *Match* criteria were introduced to see how realistic are predictions, regardless the number of sampled typologies. Thus the large geomorphologic units like lagoons (deep, intermediate and shallow), subtidal and deep subtidal reef flats are generally better predicted than the small surface typologies like Pass, Inner slope, Outer slope and Drowned subtidal reef flat typologies. Indeed, such typologies are under-represented in the study area and represent together no more than 10% of the surface. Even evaluating the predictability of such typologies was difficult using the *Match* metric, hence the use of *Balanced Match* metric pondering typologies by their surface. Using both these metrics, optimal map construction was obtained for 200 ground truth data points obtained using CDS methodology and 100 m DBM resolution (Fig. 8).

However, some limitation of the methodology used is to be noted. SCS-CLARA and CDS sampling methods are

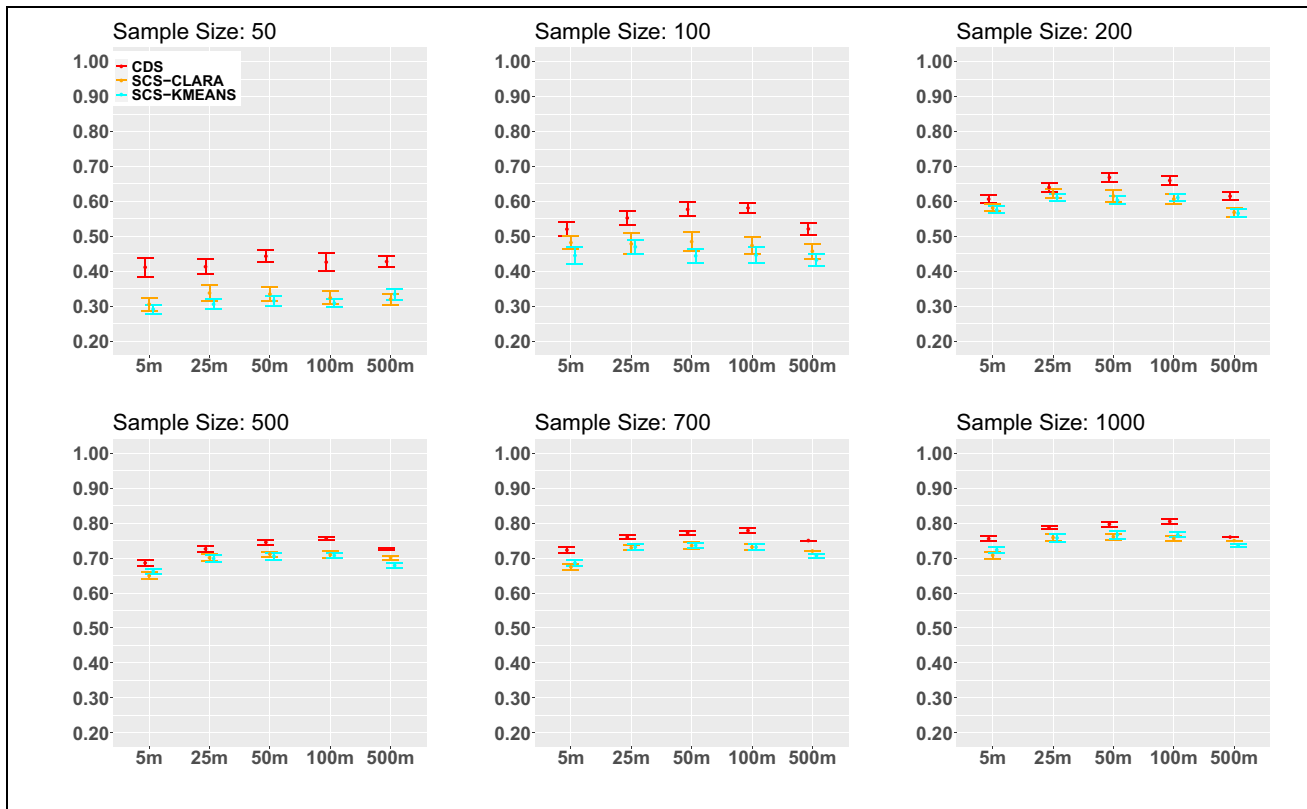


Fig. 7 Mean (+/- standard error) over 30 draws of the *Balanced Match* metric (y-axis) at different DBM resolutions (x-axis) for different sample size and different sampling methods (CDS, SCS-CLARA and SCS-KMEANS)

particularly useful for high dimensional data. The CLARA clustering algorithm can be used on large DBM grid cells data to generate ground truth locations. However, SCS-KMEANS although suitable for spatial coverage sampling, lead to computational deadlock for high DBM resolutions. Furthermore, it can not be used when there is missing data areas in the studied surface. Bathymetric imputation of poorly sampled zones was done using ordinary kriging method that performed much better than others spatial interpolation methods (cf. Table S1, an example on the supplemental materials).

Modeling choices

Traditional approaches to build geomorphologic maps are often time-consuming, labor-intensive and has a limited coverage and scale. Remote Sensing techniques although allowing wide coverage and high-resolution, require expertise in image interpretation (Gao 2009; Gilvear and Bryant 2016). GIS approaches allow for the integration and analysis of diverse data types support data visualization and complex spatial analysis. But they also require specialized software and expertise (Guzzetti et al. 1999; Napieralski and Li 2007). In addition, interpretation and analysis heavily rely on

data quality. Besides these techniques, semi-automated and automated approaches through machine learning and deep learning are increasingly used to identify complex patterns and features for landform classification, feature detection, or segmentation. Deep learning models require large amounts of labeled training data (e.g., images, point clouds) to effectively learn complex patterns and relationships. Training such models involves adjusting millions of parameters through backpropagation and optimization algorithms (e.g., stochastic gradient descent) and their tuning involves finding the right network architecture and hyperparameters. They are often considered black boxes due to their complex architectures, making it challenging to interpret the features that influence predictions (Li et al. 2020). The statistical learning approach used in this study can handle large volumes of data, automate analysis processes and discover complex patterns and relationships in the data. Interpretability of used models may also be a challenge but less than deep learning ones. In addition, it can work effectively with smaller datasets which was crucial in our objectives.

RF model for geomorphologic units clustering is chosen because RF exhibits complex and non-linear relationships between the features and the dependent variable but also on

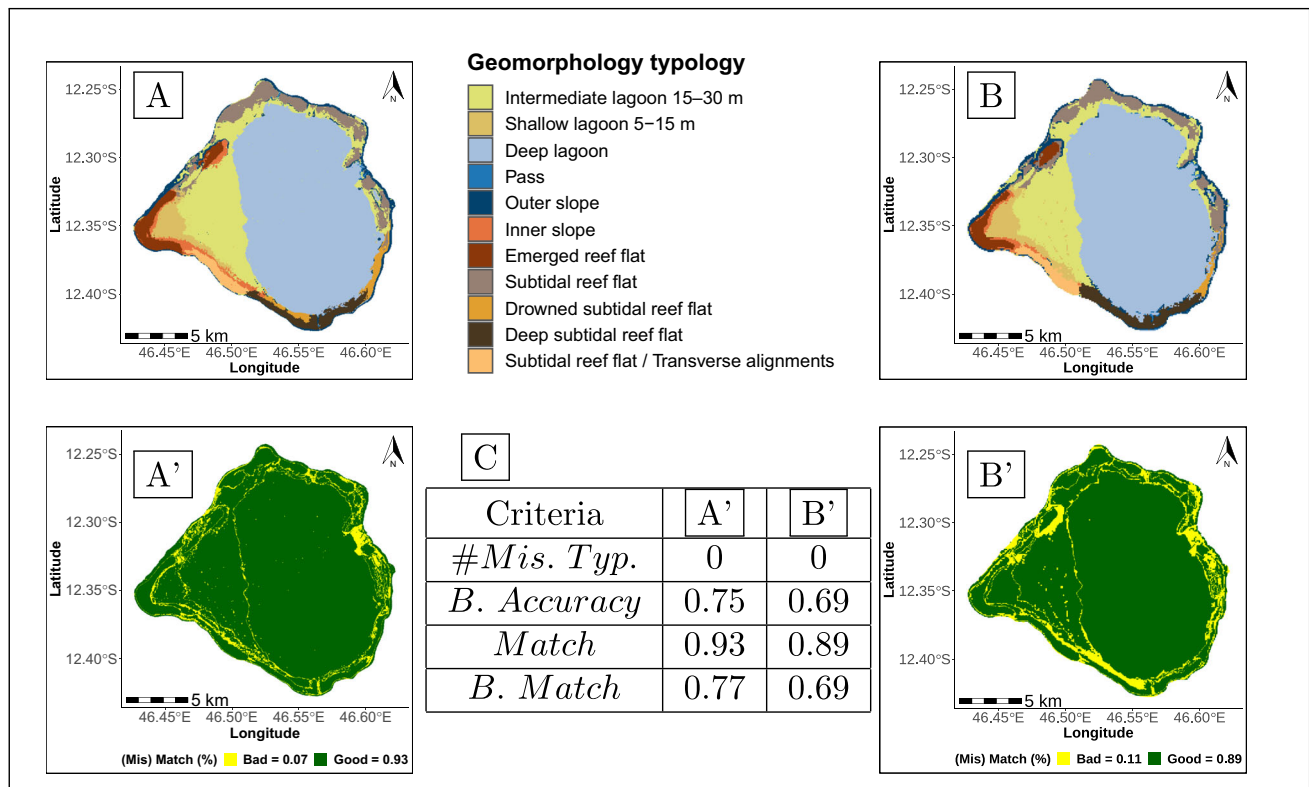


Fig. 8 **A** Predicted map with a 50 m DBM and 1000 sampled locations using CDS methodology. **B** Predicted map with a 100 m DBM and 200 sampled locations using CDS methodology. Corresponding match / mismatch maps (A' and B') in comparison to the expert one. **C** Performance criteria measured

features among themselves, handling effectively these relationships by using an ensemble of decision trees. It also provide valuable information as feature importance measure helping identify the most informative feature which made it preferable to others clustering techniques in many cases. RF was used for supervised classification problems in many recent studies, compared different algorithms has shown effectiveness of RF based algorithms (Zeraatpisheh et al. 2017; Giaccone et al. 2022). It has shown its robustness specifically when data contain uncertainties and handles high-dimensional data efficiently avoiding overfitting and reducing computational complexity.

The variable selection step in the modeling procedure is crucial in the proposed methodology. Terrain attributes calculation is not specifically time-consuming but the relevance of each of them depends on the data and the geomorphological features that are mapped. This study demonstrated that the terrain attributes, although literature based chosen initially, have not all, a great explanatory power on geomorphologic feature. For each generated sample, the number of features selected is also counted and 6 terrain attributes are generally retained.

Conclusion

A statistical learning based approach is proposed to automatically map the geomorphology of a study site using bathymetric data and some ground truth data points. On the one hand, tools to help geomorphologists to plan field campaigns in advance through an optimal DBM resolution and an automated sampling methodology to achieve field verifications are provided. On the other hand, a flexibility in the proposed methodology allowing the usage of terrain attributes as much as desired since the feature selection will help to keep only the most relevant ones is preferred. In addition, statistical and computational tools to compare geomorphological maps produced at different resolutions are provided. The methodology reproducibility is made possible by a set of reusable R scripts.

In the future, an application of the methodology using others data available in others sites is planned. An investigation on others sampling methodology for e.g. which would take into account the presence of non sampling sites inside a study area is also being considered.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12145-024-01347-x>.

Acknowledgements This work was carried out as part of a doctoral contract at the 12S doctoral school of the University of Montpellier. Geomorphological map used in this project was generated during the EPICTURE project jointly carried by IFREMER and CUFIR institutes and funded by the 10th Developmental European Funds.

We are thankful to Rodolphe Devillers and Christophe Crambes for their detailed advice and helpful suggestions. We also thank Baptiste Chapuisat, Ghislain Durif and François David Collain for their assistance using High Performance Computing techniques.

Author Contributions All authors made an equal contribution to both the research and writing of this paper. They collaborated on developing the methodology, conducting experiments, and analyzing the results. Their collective effort underscores their shared commitment to the completion of this submission.

Funding No funding was received for conducting this study.

Data Availability No datasets were generated or analysed during this project. The dataset analysed come from the EPICTURE project (<http://dx.doi.org/10.12770/21232c12-e409-4136-a24a-78c346518cfa>).

Code Availability The R code created during this work is open source and can be accessed in Zenodo repositories : <https://doi.org/10.5281/zenodo.8436795>

Declarations

Ethical Approval This manuscript has not been published, accepted for publication, or under editorial review for publication elsewhere.

Informed Consent The authors have given the informed consent to publish this article in the Journal Earth Science Informatics if accepted.

Competing interests The authors declare no competing interests.

References

- Adey W (1966) Distribution of saxicolous crustose corallines in the northwestern north atlantic. *J Phycol* 2:49–54. <https://doi.org/10.1111/j.1529-8817.1966.tb04593.x>
- Adey W, Macintyre I (1973) Crustose coralline algae: a re-evaluation in the geological sciences. *Geol Soc Am Bull* 84(3):883–904. [https://doi.org/10.1130/0016-7606\(1973\)84<883:CCAARI>2.0.CO;2](https://doi.org/10.1130/0016-7606(1973)84<883:CCAARI>2.0.CO;2)
- Ahn S, Sung H, Han H (2023) Classification of the world undersea geomorphic features from GEBCO 2020 grid data. *J Korean Geog Soc* 58(1):36–54
- Andréfouët S, Dirberg G (2006) Cartographie et inventaire du système récifal de wallis, futuna et alofi par imagerie satellitaire landsat 7 etm+ et orthophotographies aériennes à haute résolution spatiale. IRD, Centre de Nouméa et Service de L'Environnement de Wallis et Futuna
- Andréfouët S, Muller-Karger F, Robinson J, et al (2004) Global assessment of modern coral reef extent and diversity for regional science and management applications: a view from space. *Proceedings of the 10th International Coral Reef Symposium* 2:1732–1745
- Argyropoulou E, Argialas D, Nomikou P, et al. (2016) Automatic identification of submarine landforms using object-based image analysis in the area of north aegan basin. *Bull Geol Soc Greece* 50(3), 1605–1615. <https://doi.org/10.12681/bgs.11880>
- Arhant Y, Neyt X, Pizurica A (2023) A new deep learning neural network architecture for seafloor characterisation. In: *The 10th Military Sensing Symposium Proc*
- Azarafza M, Azarafza M, Akgün H, Atkinson PM et al (2023) Deep learning-based landslide susceptibility mapping. *Scientific Reports* 11(1):24112. <https://doi.org/10.3390/su14031734>
- Battistini R (1975) *Eléments de terminologie récifale indopacifique*. Station marine d'Endoume
- Behrens T, Schmidt K, MacMillan R et al (2018) Multi-scale digital soil mapping with deep learning. *Sci Rep* 8(1):15244. <https://doi.org/10.1038/s41598-018-33516-64>
- Biau G, Scornet E (2016) A random forest guided tour. *Test* 25:197–227. <https://doi.org/10.48550/arXiv.1511.05741>
- Bishop M, James L, Shroder J Jr et al (2012) Geospatial technologies and digital geomorphological mapping: concepts, issues and research. *Geomorphology* 137(1):5–26. <https://doi.org/10.1016/j.geomorph.2011.06.027>
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140. <https://doi.org/10.1007/BF00058655>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Breyer G, Pesch Bartholomä A, R, (2023) The suitability of machine-learning algorithms for the automatic acoustic seafloor classification of hard substrate habitats in the German bight. *Remote Sensing* 15(16):4113. <https://doi.org/10.3390/rs15164113>
- Browne N, Smithers S, Perry C (2010) Geomorphology and community structure of middle reef, central great barrier reef, Australia: an inner-shelf turbid zone reef subject to episodic mortality events. *Coral Reefs* 29:683–689. <https://doi.org/10.1007/s00338-010-0640-3>
- Brus D (2019) Sampling for digital soil mapping: a tutorial supported by R scripts. *Geoderma* 338:464–480. <https://doi.org/10.1016/j.geoderma.2018.07.036>
- Brus D, De Gruijter J, Van Groenigen J (2006) Designing spatial coverage samples using the k-means clustering algorithm. *Dev Soil Sci* 31:183–192. [https://doi.org/10.1016/S0166-2481\(06\)31014-8](https://doi.org/10.1016/S0166-2481(06)31014-8)
- Copeland A, Edinger E, Devillers R et al (2013) Marine habitat mapping in support of marine protected area management in a subarctic fjord: Gilbert Bay, Labrador, Canada. *J Coast Conserv* 17:225–237. <https://doi.org/10.1007/s11852-011-0172-1>
- Cressie N (1988) Spatial prediction and ordinary kriging. *Math Geol* 20:405–421. <https://doi.org/10.1007/BF00892986>
- Cui X, Liu H, Fan M, et al (2021) Seafloor habitat mapping using multi-beam bathymetric and backscatter intensity multi-features SVM classification framework. *Appl Acoust* 174:107728. <http://dx.doi.org/10.1016/j.apacoust.2020.107728>
- Curie F, Gaillard S, Ducharme A et al (2007) Geomorphological methods to characterise wetlands at the scale of the seine watershed. *Sci Total Environ* 75(1–3):59–68. <https://doi.org/10.1016/j.scitotenv.2006.12.013>
- Dartnell P (2000) Applying remote sensing techniques to map seafloor geology/habitat relationships. Masters Thesis, San Francisco State University
- Dekavalla M, Argialas D (2017) Object-based classification of global undersea topography and geomorphological features from the SRTM30 PLUS data. *Geomorphology* 288:66–82. <https://doi.org/10.1016/j.geomorph.2017.03.026>
- Díaz-Urriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:1–13. <https://doi.org/10.1186/1471-2105-7-3>
- Diesing M, Green S, Stephens D et al (2014) Mapping seabed sediments: comparison of manual, geostatistical, object-based image analysis and machine learning approaches. *Cont Shelf Res* 84:107–119. <https://doi.org/10.1016/j.csr.2014.05.004>
- Dong Y, Liu Y, Hu C et al (2019) Coral reef geomorphology of the Spratly Islands: a simple method based on time-series of Landsat

- 8 multi-band inundation maps. *ISPRS J Photogramm Remote Sens* 157:137–154. <https://doi.org/10.1016/j.isprsjprs.2019.09.011>
- Dramis F, Guida D, Cestari A (2011) Nature and aims of geomorphological mapping. *Dev Earth Surf Process* 15:39–73. <https://doi.org/10.1016/B978-0-444-53446-0.00003-3>
- Evans I (1980) An integrated system of terrain analysis and slope mapping. *Zeitschrift fur Geomorphologic Suppl-Bd* 36:274–295
- Florinsky I (1998) Accuracy of local topographic variables derived from digital elevation models. *Int J Geog Infor Sci* 12(1):47–61. <https://doi.org/10.1080/136588198242003>
- Fukunaga A, Craig B, Kosaki R (2019) Integrating three-dimensional benthic habitat characterization techniques into ecological monitoring of coral reefs. *J Marine Sci Eng* 7(2). <https://doi.org/10.3390/jmse7020027>
- Galvez D, Papenmeier S, Sander L et al (2022) Ensemble mapping as an alternative to baseline seafloor sediment mapping and monitoring. *Geo-Mar Lett* 42(3):11. <https://doi.org/10.1007/s00367-022-00734-x>
- Gao J (2009) Bathymetric mapping by means of remote sensing: methods, accuracy and limitations. *Prog Phys Geogr* 33(1):103–116. <https://doi.org/10.1177/0309133309105657>
- Giaccone E, Oriani F, Tonini M et al (2022) Using data-driven algorithms for semi-automated geomorphological mapping. *Stoch Environ Res Risk Assess* 36:2115–2131. <https://doi.org/10.1007/s00477-021-02062-5>
- Gilvear D, Bryant R (2016) Analysis of remotely sensed data for fluvial geomorphology and river science. Tools in fluvial geomorphology pp 103–132. <https://doi.org/10.1002/9781118648551.ch6>
- Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. [arXiv:2008.05756](https://arxiv.org/abs/2008.05756) <https://doi.org/10.48550/arXiv.2008.05756>
- Guyon I, Elisseeff A (2020) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182. <https://doi.org/10.1162/153244303322753616>
- Guyon I, Weston J, Barnhill S (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46:389–422. <https://doi.org/10.1023/A:1012487302797>
- Guzzetti F, Carrara A, Cardinali M et al (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology* 31(1–4):181–216. [https://doi.org/10.1016/S0169-555X\(99\)00078-1](https://doi.org/10.1016/S0169-555X(99)00078-1)
- Hartigan J (1975) Clustering algorithms. John Wiley & Sons Inc
- Horn B (1981) Hill shading and the reflectance map. *Proc IEEE* 69(1):14–47. <https://doi.org/10.1109/PROC.1981.11918>
- Hugenholtz C, Whitehead B, Brown O et al (2013) Geomorphological mapping with a small unmanned aircraft system (sUAS): feature detection and accuracy assessment of a photogrammetrically-derived digital terrain model. *Geomorphology* 194:16–24. <https://doi.org/10.1016/j.geomorph.2013.03.023>
- Ilich A, Misiuk B, Lecours V et al (2023) MultiscaleDTM: An open-source R package for multiscale geomorphometric analysis. *Trans GIS* 4:1164–1204. <https://doi.org/10.1111/tgis.13067>
- Janowski L, Wroblewski R, Rucinska M, Kubowicz-Grajewska A et al (2022) Automatic classification and mapping of the seabed using airborne LiDAR bathymetry. *Eng Geol* 301:106615. <https://doi.org/10.1016/j.enggeo.2022.106615>
- Janowski L, Wroblewski R, Dworniczak J et al (2021) Offshore benthic habitat mapping based on object-based image analysis and geomorphometric approach. A case study from the Slupsk Bank, Southern Baltic sea. *Sci Total Env* 801:149712. <https://doi.org/10.1016/j.scitotenv.2021.149712>
- Jasiewicz J, Stepinski T (2013) Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology* 182:147–156. <https://doi.org/10.1016/j.geomorph.2012.11.005>
- Kaufman L, Rousseeuw P (1975) Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis 344:68–125
- Kienholz H (1978) Maps of geomorphology and natural hazards of Grindelwald, Switzerland: Scale 1: 10,000. *Arct Alp Res* 10(2):169–184
- Koop L, Snellen M, Simons D (2021) An object-based image analysis approach using bathymetry and bathymetric derivatives to classify the seafloor. *Geosciences* 11:45. <https://doi.org/10.3390/geosciences11020045>
- Kuhn M (2019) caret: classification and regression training. R package, version 6.0-92. Accessed: 2023 Mar 28. <https://CRAN.R-project.org/package=caret>
- Lacharité M, Brown C, Gazzola V (2018) Multisource multibeam backscatter data: developing a strategy for the production of benthic habitat maps using semi-automated seafloor classification methods. *Mar Geophys Res* 39:307–322. <https://doi.org/10.1007/s11001-017-9331-6>
- Lecours V, Dolan M, Micallef A et al (2016) A review of marine geomorphometry, the quantitative study of the seafloor. *Hydrol Earth Syst Sci* 20(8):3207–3244. <https://doi.org/10.5194/hess-20-3207-2016>
- Leon J, Roelfsema C, Saunders M et al (2015) Measuring coral reef terrain roughness using “structure-from-motion” close-range photogrammetry. *Geomorphology* 242:21–28. <https://doi.org/10.1016/j.geomorph.2015.01.030>
- Li S, Xiong L, Tang G et al (2020) Deep learning-based approach for landform classification from integrated data sources of digital elevation model and imagery. *Geomorphology* 354:107045. <https://doi.org/10.1016/j.geomorph.2020.107045>
- Locker S, Armstrong R, Battista T et al (2010) Geomorphology of mesophotic coral ecosystems: current perspectives on morphology, distribution, and mapping strategies. *Coral Reefs* 29:329–345. <https://doi.org/10.1007/s00338-010-0613-6>
- Lucieer V, Lucieer A (2009) Fuzzy clustering for seafloor classification. *Mar Geol* 264(3–4):230–241. <https://doi.org/10.1016/j.margeo.2009.06.006>
- Lundblad E, Wright D, Miller J et al (2006) A benthic terrain classification scheme for American Samoa. *Mar Geodesy* 29(2):89–111. <https://doi.org/10.1080/01490410600738021>
- Maschmeyer C, White S, Dreyer B et al (2019) High-silica lava morphology at ocean spreading ridges: machine-learning seafloor classification at Alarcon Rise. *Geosciences* 9(6):245. <https://doi.org/10.3390/geosciences9060245>
- Masetti G, Mayer L, Ward L (2018) A bathymetry-and reflectivity-based approach for seafloor segmentation. *Geosciences* 8(1):14. <https://doi.org/10.3390/geosciences8010014>
- Mata D, Úbeda J, Fernández-Sánchez A (2021) Modelling of the reef benthic habitat distribution within the Cabrera National Park (Western Mediterranean Sea). *Ann GIS* 27(3):285–298. <https://doi.org/10.1080/19475683.2021.1936169>
- Minár J, Evans I (2008) Elementary forms for land surface segmentation: the theoretical basis of terrain analysis and geomorphological mapping. *Geomorphology* 95(3–4):236–259. <https://doi.org/10.1016/j.geomorph.2007.06.003>
- Minnery G, Rezak R, Bright T (1985) Depth zonation and growth form of crustose coralline algae: flower garden banks, northwestern gulf of Mexico. *Paleoalgology: Contemporary research and applications* Berlin, Heidelberg: Springer p 237–246. https://doi.org/10.1007/978-3-642-70355-3_18
- Misiuk B, Brown C (2023) Improved environmental mapping and validation using bagging models with spatially clustered data. *Eco Inform* 77:102181. <https://doi.org/10.1016/j.ecoinf.2023.102181>
- Misiuk B, Diesing M, Aitken A et al (2021) A spatially explicit comparison of quantitative and categorical modelling approaches

- for mapping seabed sediments using random forest. *Ann GIS* 9(6):254. <https://doi.org/10.3390/geosciences9060254>
- Napieralski J, Harbor J, Li Y (2007) Glacial geomorphology and geographic information systems. *Earth Sci Rev* 85(1–2):1–22. <https://doi.org/10.1016/j.earscirev.2007.06.003>
- Novaczek E, Devillers R, Edinger E (2019) Generating higher resolution regional seafloor maps from crowd-sourced bathymetry. *PLoS ONE* 14(6):e0216792. <https://doi.org/10.1371/journal.pone.0216792>
- Oshiro T, Perez P, Baranauskas J (2012) How many trees in a random forest? *Machine Learning and Data Mining in Pattern Recognition MLDM 2012 Lecture Notes in Computer Science()*. Springer, Berlin, Heidelberg, pp 7376. https://doi.org/10.1007/978-3-642-31537-4_13
- Otto JC, Prasicsek G, Blöthe J, et al (2018) GIS applications in geomorphology. In: *Comprehensive geographic information systems*. Elsevier, p 81–111, <https://doi.org/10.1016/B978-0-12-409548-9.10029-6>
- Otto JC, Smith M (2013) *Geomorphological mapping*, vol Section 2.6. *British Soc Geomorphol chap 2:1–10*
- Pandian P, Ruscoe J, Shields M, et al (2009) Seabed habitat mapping techniques: an overview of the performance of various systems. *Med Marine Sci* 10(2):29–44. <https://doi.org/10.12681/mms.107>
- Pavlopoulos K, Evelpidou N, Vassilopoulos A (2009) *Mapping geomorphological environments*. Springer Science & Business Media. <https://doi.org/10.1007/978-3-642-01950-0>
- Probst P, Wright M, Boulesteix A (2019) Hyperparameters and tuning strategies for random forest. *Wiley Interdis Rev Data Mining know Disc* 9(3):e1301. <https://doi.org/10.1002/widm.1301>
- Riley S, DeGloria S, Elliot R (1999) Index that quantifies topographic heterogeneity. *Int J Sci* 5(1–4):23–27
- Roberts D (2015) Spatially balanced subsampling in r (retaining maximum sample size). <https://davidrroberts.wordpress.com/2015/09/25/spatial-buffering-of-points-in-r-while-retaining-maximum-sample-size/>
- Roos D, Dupont P, Gaboriau M, et al (2017) Projet épique : Etude des peuplements ichthyologiques et des communautés récifales à partir d'indicateurs spatiaux et de l'approche fonctionnelle, des bancs du geyser, de la zélée et de l'iris. <https://doi.org/10.13155/54549>
- Royle J, Nychka D (1998) An algorithm for the construction of spatial coverage designs with implementation in spls. *Comput Geosci* 24(5):479–488
- Rozycka M, Migon P, Michniewicz A (2017) Topographic wetness index and terrain ruggedness index in geomorphic characterisation of landslide terrains, on examples from the sudetes, sw poland. *Zeitschrift für Geomorphologie Supplementary issues* 61(2):61–80
- Sappington J, Longshore K, Thompson D (2007) Quantifying landscape ruggedness for animal habitat analysis: A case study using bighorn sheep in the Mojave Desert. *J Wildl Manag* 71(5):1419–1426. <https://doi.org/10.2193/2005-723>
- Schmidt J, Hewitt A (2004) Fuzzy land element classification from DTMs based on geometry and terrain position. *Geoderma* 121(3–4):243–256. <https://doi.org/10.1016/j.geoderma.2003.10.008>
- Siart C, Bubbenzer O, Eitel B (2009) Combining digital elevation data (SRTM/ASTER), high resolution satellite imagery (Quickbird) and GIS for geomorphological mapping: a multi-component case study on Mediterranean karst in Central Crete. *Geomorphology* 112(1–2):106–121. <https://doi.org/10.1016/j.geomorph.2009.05.010>
- Siqueira R, Veloso G, Fernandes-Filho E et al (2022) Evaluation of machine learning algorithms to classify and map landforms in Antarctica. *Earth Surf Proc Land* 47(2):367–382. <https://doi.org/10.1002/esp.5253>
- Sklar E, Bushuev E, Misiuk B et al (2024) Seafloor morphology and substrate mapping in the Gulf of St Lawrence, Canada, using machine learning approaches. *Front Mar Sci* 11:1306396. <https://doi.org/10.3389/fmars.2024.1306396>
- Sowers D, Masetti G, Mayer L et al (2020) Standardized geomorphic classification of seafloor within the United States Atlantic canyons and continental margin. *Front Mar Sci* 7:9. <https://doi.org/10.3389/fmars.2020.00009>
- Stepinski T, Ghosh S, Vilalta R (2007) Machine learning for automatic mapping of planetary surfaces. *Aqua Conserv Marine Freshwater Ecosyst* 30(4):846–859. <https://doi.org/10.13140/2.1.1518.9445>
- Sterne T, Retchless D, Allee R et al (2020) Predictive modelling of mesophotic habitats in the north-western Gulf of Mexico. *Proc Natl Conf Artif Intell* 22(2):1807. <https://doi.org/10.1002/aqc.3281>
- Strobl C, Boulesteix A, Kneib T et al (2008) Conditional variable importance for random forests. *BMC Bioinformatics* 9:1–11. <https://doi.org/10.1002/aqc.3281>
- Summers G, Lim A, Wheeler A (2021) A scalable, supervised classification of seabed sediment waves using an object-based image analysis approach. *Remote Sensing* 13(12):2317. <https://doi.org/10.3390/rs13122317>
- Valentine P, Fuller S, Scully L (2004) Terrain ruggedness analysis and distribution of boulder ridges in the stellwagen bank national marine sanctuary region (poster). Galway, Ireland: 5th International Symposium on Marine Geological and Biological Habitat Mapping (GeoHAB)
- Van der Meij W, Meijles E, Marcos D et al (2022) Comparing geomorphological maps made manually and by deep learning. *Earth Surf Proc Land* 47(4):1089–1107. <https://doi.org/10.1002/esp.5305>
- Wabnitz C, Andréfouët S, Torres-Pulliza D et al (2008) Regional-scale seagrass habitat mapping in the Wider Caribbean region using Landsat sensors: applications to conservation and ecology. *Remote Sens Environ* 112(8):3455–3467. <https://doi.org/10.1016/j.rse.2008.01.020>
- Wilson M, O'Connell B, Brown C et al (2007) Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Marine Geodesy* 30:3–35. <https://doi.org/10.1080/01490410701295962>
- Wynn R, Huvenne V, Le Bas T et al (2014) Autonomous underwater vehicles (AUVs): their past, present and future contributions to the advancement of marine geoscience. *Mar Geol* 352:451–468. <https://doi.org/10.1016/j.margeo.2014.03.012>
- Zeraatpisheh M, Ayoubi S, Jafari A et al (2017) Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology* 285:186–204. <https://doi.org/10.1016/j.geomorph.2017.02.015>
- Zevenbergen L, Thorne C (1987) Quantitative analysis of land surface topography. *Earth Surf Proc Land* 12:47–56. <https://doi.org/10.1002/esp.3290120107>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Paul Aimé Latsouck Faye¹ · Elodie Brunel¹ · Thomas Claverie^{2,3} · Solym Mawaki Manou-Abi^{1,3,4} · Sophie Dabo-Niang⁵

✉ Paul Aimé Latsouck Faye
paul-aime-latsouck.faye@umontpellier.fr

Elodie Brunel
elodie.brunel-piccinini@umontpellier.fr

Thomas Claverie
thomas.claverie@univ-reunion.fr

Solym Mawaki Manou-Abi
solym.manou-abi@univ-mayotte.fr

Sophie Dabo-Niang
sophie.dabo@univ-lille.fr

¹ IMAG, University of Montpellier, CNRS, 34090 Montpellier, France

² UMR ENTROPIE, IRD, IFREMER, CNRS, Univ La Réunion, Saint Denis 97744, Réunion, France

³ Université de Mayotte, Dembèni 97660, Mayotte, France

⁴ Laboratoire de Mathématiques et Applications UMR 7348, University of Poitiers, CNRS, Futuroscope, 86073 Poitiers, France

⁵ PAINLEVE UMR 8524, University of Lille, CNRS, INRIA-MODAL, 59665 Lille, France

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com



Robust estimator of the ruin probability in infinite time for heavy-tailed distributions

El Hadji Deme^a, Yousri Slaoui^b, Modou Kebe^a and Solym Manou-Abi^{c,d}

^aLERSTAD, UFR SAT, Université Gaston Berger, Saint-Louis, Sénégal; ^bLaboratoire de Mathématiques et Applications, UMR CNRS 7348, Université de Poitiers, Poitiers, France; ^cInstitut Montpellierain Alexander Grothendieck, UMR CNRS 5149, Montpellier, France; ^dCentre Universitaire de Formation et de Recherche, Mayotte, France

ABSTRACT

The probability of ruin of an insurance company is one of the main risk measures considered in risk theory, and the problems of its calculation and approximation have attracted much attention. Statistical estimations have been developed on the ruin probability in infinite time for insurance losses from heavy-tailed distributions. However, these estimations suffer heavily from under-coverage or have a robustness problem. We therefore need another method for estimating the probability of ruin in infinite time for heavy-tailed losses. In this paper, we introduce a robust estimator of the infinite-time probability of ruin for such distributions. Our methodology is based on extreme value theory, which offers adequate statistical results for such distributions. Our approach is based on a sensitive distribution known as the t-Hill estimator (t-score or score moment estimation) for the index of any tail distribution and introduced in Fabián and Stehlík. We establish their asymptotic normality, and through a simulation study, illustrate their behaviour in terms of absolute bias and mean squared error. The simulation results show that our estimators perform well and that they are fairly robust to outliers.

ARTICLE HISTORY

Received 8 July 2023
Accepted 4 September 2024

KEYWORDS

Ruin probability; estimation; robustness; tail index; heavy-tailed; reinsurance

2020 MATHEMATICS

SUBJECT

CLASSIFICATIONS

62H12; 62H05; 60G70

1. Introduction

The effect of insurance operations is the total or partial transfer of the financial consequences of the risk incurred by the insured to an insurance company. The expenses covered by the company may correspond either to indemnities to be paid to third parties in respect of the insured's liability (civil, professional or other), or to compensation for damage suffered by the insured. But what happens to these insurance companies when they themselves are exposed to risk? One approach to this problem is based on the use of ruin theory (see, e.g., [1]).

In the field of insurance, risk is defined as the probability that an insurance company's reserve, i.e., the difference between the premium amount and the insurance amount, will be reduced of an insurance company, which is the difference between the total premiums

received and the total amount of claims paid, becomes negative at some point. At that point, ruin is said to occur, due to a miscalculation of the policyholders' contribution rate, or claims that are too large to cover. Indeed, the probability of such an event is seen as a means of controlling risk behaviour. It is also a useful way of controlling the insurer's funds in long-term planning. Let's recall the definition of a standard mathematical model for insurance risk (see, for example, [2], p. 345).

The insurance company's initial capital is denoted u . The number of claims over the period $(0, t]$, denoted M_t , is described by a Poisson process with a fixed intensity (rate) $\lambda > 0$. We also define a sequence of non-negative random variables $\{X_j\}_{j=1}^{\infty}$ independent and identically distributed (i.i.d.) with a loss distribution function (df) $F_1(x) := \mathbb{P}(X_j \leq x)$ representing loss severity, with unknown finite mean $\mu_1 := \mathbb{E}(X_1) = \int_0^{\infty} (1 - F_1(x)) dx$. The finiteness of the mean of the losses X_j , $j = 1, \dots, n$ ensures that its corresponding excess of losses $X_j - u$, $X_j > u$, $j = 1; \dots, n$ has also a finite mean. In addition, reinsurance companies have to calculate premiums to cover these excess claims, which are generally very high and lead to infinite second moments of the loss variables

The known of this mean value has been of great interest to insurance companies, since it is one of the most commonly used premium calculation principles, known as *the net premium*, and corresponds to the expected amount of claims for a given insurance period.

Let's then assume that X_j 's are independent of M_t and that the insurer collects a premium at a constant rate c per unit time and that the net profit condition is met, i.e., $c/\lambda > \mu_1$. The classical risk process $\{R_t\}_{t>0}$ is given by:

$$R_t := u + ct - \sum_{j=1}^{M_t} X_j, \quad t > 0.$$

The corresponding claim surplus process is defined by

$$S_t := u - R_t = ct - \sum_{j=1}^{M_t} X_j, \quad t > 0.$$

First of all, we're interested in the probability that S_t exceeds an initial reserve u at a time t before or at a horizon T . Explicitly, this probability can be written as follows:

$$\psi_T(u) := \mathbb{P} \left\{ \sup_{0 < t \leq T} S_t > u \right\}.$$

The ruin probability in infinite time is defined by,

$$\phi(u) := \lim_{T \rightarrow \infty} \psi_T(u). \quad (1)$$

In the actuarial field, the costs of large claims require the modelling of rare events, i.e., events with a low probability of occurrence, but with large claims amounts and disastrous effects. The analysis of these extreme events can be carried out using the extreme value methodology, whose distribution functions F_1 are heavy-tailed and mainly characterized by their index, which indicates the size and frequency of certain extreme phenomena within a given probability distribution (see for example [3]).

The heavy-tailed nature of claims requires particular attention to the analysis of tail distributions. Extreme value theory (EVT) therefore offers suitable statistical tools for modelling these distribution tails, see for example [3–9], among many others.

The concern of reinsurance companies is to be interested in premium calculations to cover these excess losses, which are generally very high values. The extreme value theory has become one of the leading theories in the development of statistical models for such reinsurance losses.

Now, suppose that F_1 is heavy-tailed, that is:

$$\lim_{x \rightarrow \infty} \exp(\delta x) (1 - F_1) = \infty, \quad \text{for all } \delta > 0. \tag{2}$$

The class of regularly varying functions provides good examples of heavy-tailed models. We can cite the following models: Pareto, Burr, Student, Lévy-stable and log-gamma (see, for example, [10]). In the remainder of this paper, we restrict ourselves to this class of distributions. In other words, we assume that the survival function $1 - F_1$ is regularly varying at infinity with index $-1/\gamma < 0$, that

$$\lim_{z \rightarrow \infty} \frac{1 - F_1(xz)}{1 - F_1(z)} = x^{-1/\gamma}, \quad \text{for any } x > 0. \tag{3}$$

The parameter γ is the tail index and governs tail behaviour, with higher values indicating heavier tails. For more details on these models, we can refer to [9,11,12]. It has been shown that for large initial reserve u , the ruin probability $\phi(u)$ can be approximated, under the assumption, by:

$$\phi(u) \sim \left(\frac{c}{\lambda} - \mu_1\right)^{-1} \int_u^\infty (1 - F_1(x)) dx, \tag{4}$$

(see, e.g., [13]). A change of variables yields to $\int_u^\infty (1 - F_1(x)) dx = \int_0^\infty (1 - F_1(z + u)) dz$. Now, let F_2 be the distribution function of the stop-loss variables $Y_i := \max(X_j - u)_+ = \max(X_j - u, 0)$, $j = 1; \dots, n$, we have $F_2(z) = F_1(z + u)$, for $z > 0$ and $F_2(z) = 0$, otherwise. Since the mean of loss variable Y_1 is given by $\mu_2 := \mathbb{E}(Y_1) = \int_0^\infty (1 - F_2(z)) dz$ and is finite under the finiteness of μ_1 , then denoted by $\omega := c/\lambda$, the approximation in (4) can be rewritten as:

$$\phi(u) \sim \frac{\mu_2}{\omega - \mu_1}. \tag{5}$$

Also, for a fixed initial reserve u and for all $z > 0$ and for any $x > 0$:

$$\frac{1 - F_2(xz)}{1 - F_2(z)} = \frac{1 - F_1(xz + u)}{1 - F_1(z + u)}. \tag{6}$$

Since the function $z \mapsto z + u$ is regularly varying at infinity with index $\beta = 1$, and by remarking that $z + u \rightarrow \infty$, as $t \rightarrow \infty$, then under the assumption (3), we have from the properties of regularly wearying functions in the 2nd assertion of Proposition B.1.9, p. 366 in [14], $1 - F(\cdot + u)$ is regularly varying at infinity with index $-\beta/\gamma = -1/\gamma$ and

$$\lim_{z \rightarrow \infty} \frac{1 - F_2(xz)}{1 - F_2(z)} = \lim_{z \rightarrow \infty} \frac{1 - F_1(xz + u)}{1 - F_1(z + u)} = x^{-1/\gamma} \quad \text{for any } x > 0. \tag{7}$$

This means that the survival function $1 - F_2$ of the stop loss variable Y is regularly varying at infinity with index $-1/\gamma$, as $1 - F_1$ is.

Now, let Q_i , $i \in \{1, 2\}$ be the generalized inverse functions (or quantile functions) related to df F_i , $i \in \{1, 2\}$ and defined as follows for all $s \in (0, 1]$:

$$Q_i(s) := \inf \{x > 0 : F_i(x) \geq s\}.$$

From Corollary 1.2.10 (p. 23) in [14], we have for any $x > 0$

$$\lim_{s \downarrow 0} \frac{Q_i(1 - sx)}{Q_i(1 - s)} = x^{-\gamma}, \quad i \in \{1, 2\}. \quad (8)$$

By a change of variable, the expected values $\mu_i = \int_0^\infty (1 - F_i(x)) dx$, $i \in \{1, 2\}$ can be approximated in terms of quantile function Q_i as follows: $\mu_i = \int_0^1 Q_i(s) ds$, $i \in \{1, 2\}$. Thus, the probability of ruin in infinite time is equal to:

$$\phi(u) \sim \frac{\int_0^1 Q_2(s) ds}{\omega - \int_0^1 Q_1(s) ds}. \quad (9)$$

Next, we note that:

- When $\gamma > 1$, the first moments $\mu_1 = \int_0^1 Q_1(s) ds$ and $\mu_2 = \int_0^1 Q_2(s) ds$ of losses are not defined. Thus, their associated ruin probability $\phi(u)$ is also not defined and it is not possible to do its statistical estimation.
- When $0 < \gamma \leq 1/2$ (the lower half of the unit interval), then $\mathbb{E}[X_1^{2+\epsilon}] < \infty$ and $\mathbb{E}[Y_1^{2+\epsilon}] < \infty$ for some $\epsilon > 0$, and so a nonparametric estimator for the ruin probability $\phi(u)$ can be obtain by the ratio of two empirical means and from [15], this estimator is asymptotically normal.
- When $1/2 < \gamma < 1$ (the upper half of the unit interval), then the second moment is infinite, and so the asymptotic normality of the nonparametric estimator of $\phi(u)$. is violated.

The last situation motivate the need of a specific estimator of the ruin probability for heavy-tailed loss distributions with infinite second moments, that is with index in the upper half of the unit interval ($1/2 < \gamma < 1$). The class of heavy-tailed distributions (the so-called Pareto-type distributions) includes distributions such as Pareto, Burr, Student, Lévy-stable, and log-gamma which are known to be appropriate models in Extreme Value Theory for fitting large insurance claims, large fluctuations of prices, log-returns, (see, e.g., [4,5,9,16,17]).

Statistical estimation has been developed by Rassoul [18] on the ruin probability in infinite time for insurance losses from heavy-tailed distributions with index in the upper half of the unit interval. In its framework, the author used a classical extreme value index estimator introduced by Hill [19].

Unfortunately, the Hill's estimator suffers heavily of robustness particularly when losses are contaminated by large variations in the arrival of claims and large values. This makes that estimator of the ruin probability at infinite time sensitive and constitutes a serious problem. The same problem has been observed by Brahim and Kenioua [20] and Bouali et al. [21], who have been introduced Robust estimators of actuarial risk measures for heavy-tailed losses.

In their frameworks, these authors have been solved the aforementioned problem of the classical Hill estimator, by substituting it in the estimation of risk premiums with a robust estimator (the so-called t-Hill estimator) of the extreme value index introduced in [22]. Also, Jordanova et al. [23] noted that the t-Hill estimator is better than the Hill estimator in case when the expectation does not exist, and consider not only the case of independent identically distributed observations, but also the case when the data come from moving average time series.

In this work, we introduce a robust estimator of the infinite-time probability of ruin for heavy-tailed distributions with infinite second moments. Our consideration is based on a sensitive distribution known as the t-Hill estimator (t-score or score moment estimation) for the extreme value index, which solve the robustness of the estimator proposed in [18].

The rest of the paper is organized as follows. In Section 2, we present some preliminaries on classical estimators of the probability of ruin $\phi(u)$. As these estimates suffer greatly from under-coverage or have a robustness problem, especially when losses are contaminated by large variations in the arrival of claims, we introduce, in Subsection 3.1, a robust infinite-time estimator of the probability of ruin for heavy-tailed insured losses. Using extreme value methodology, we establish its asymptotic distribution in Subsection 3.2. In Subsection 4.1, we perform a simulation study to illustrate the behaviour of our robust estimator compared to the classical estimator in terms of absolute bias and mean squared error. In Subsection 4.2, we present a contamination study in which the robustness of the estimator is evaluated. All proofs are reported in Section 5.

2. Estimating the ruin probability $\phi(u)$

First, we set a large initial reserve u . Let (X_1, \dots, X_n) and (Y_1, \dots, Y_n) be two independent samples of risks X and Y respectively. The non-parametric estimators of the distribution functions F_1 and F_2 are respectively defined as follows $F_{1,n}(x) = n^{-1} \sum_{j=1}^n \mathbb{I}_{\{X_j \leq x\}}$ and $F_{2,n}(y) = n^{-1} \sum_{j=1}^n \mathbb{I}_{\{Y_j \leq y\}}$. Thus, their corresponding empirical quantile functions are expressed by $Q_{i,n}(s) = \inf\{x; F_{i,n}(x) \geq s\}, i \in \{1, 2\}$ where \mathbb{I}_S is the indicator function of the set S . Let's denote by $X_{1,n} \leq \dots \leq X_{n,n}$ and $Y_{1,n} \leq \dots \leq Y_{n,n}$ the order statistics associated respectively with the samples (X_1, \dots, X_n) and (Y_1, \dots, Y_n) . Therefore, $Q_{1,n}(t) = X_{j,n}$ and $Q_{2,n}(t) = Y_{j,n}$ for all $t \in ((j - 1)/n, j/n]$, and for all $j = 1, \dots, n$.

To this end, a natural candidate for the empirical estimator of $\phi(u)$ is obtained by replacing in (9) the real quantiles $Q_1(\cdot)$ and $Q_2(\cdot)$ by their respective sample quantiles $Q_{1,n}(\cdot)$ and $Q_{2,n}(\cdot)$. The 'traditional' non-parametric estimator of the probability of ruin, can be approximated as follows:

$$\bar{\phi}_n(u) \sim \frac{\bar{Y}}{\omega - \bar{X}}. \tag{10}$$

where $\bar{X} := \frac{1}{n} \sum_{j=1}^n X_j$ and $\bar{Y} := \frac{1}{n} \sum_{j=1}^n Y_j$ are the sample estimators the mean of μ_1 and μ_2 respectively.

Note that for $\gamma \geq 1$, the theoretical mean of X (respectively Y) does not exist. Consequently, we will focus exclusively on distributions whose tail indices lie in the unit interval $0 < \gamma < 1$.

Next, the random variable $\sqrt{n}(\bar{\phi}_n(u) - \phi(u))$ is asymptotically equivalent to :

$$\frac{1}{(\omega - \bar{X})(\omega - \mu_1)} \sqrt{n} ((\bar{Y} - \mu_1)(\omega - \mu_1) + (\bar{X} - \mu_1)\mu_2).$$

According to the law of large numbers (LLN), the random variable $(\omega - \bar{X})$ converges in probability to $(\omega - \mu_1)$. Using asymptotic theory for L-statistics (e.g., Shorack and Wellner [24]), and the underlying distributions with a sufficient number of finite moments, we obtain from [15] the following asymptotic normality result:

$$\sqrt{n}(\bar{\phi}_n(u) - \phi(u)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_\phi^2), \quad \text{as } n \rightarrow \infty, \quad (11)$$

where

$$\sigma_\phi^2 = \frac{1}{(\omega - \mu_1)^4} ((\omega - \mu_1)^2 \sigma_2^2 + \mu_2^2 \sigma_1^2 + \mu_2(\omega - \mu_1) \sigma_{1,2}) < \infty,$$

with σ_1^2 and σ_2^2 represent respectively the variances of X and Y and

$$\sigma_{1,2} = \int_0^1 \int_0^1 (\min(s, t) - st) dQ_1(1-s) dQ_2(1-t).$$

Note that in the case where the tail index γ is in the lower half of the unit interval, i.e., $0 < \gamma < \frac{1}{2}$, the second-order moments for the two random variables X and Y are finite. Consequently, the asymptotic normality of $\phi_n(u)$ in (11) holds. This result is not respected when the loss distribution is heavy-tailed with an index γ located in the upper half of the unit interval, i.e., $1/2 < \gamma < 1$, since the asymptotic variance σ_ϕ^2 is infinite, which is due in this case to the inevitability of the second-order moments of the loss X . In this case, $\phi(u)$ must be estimated using another approach that would guarantee asymptotic normality.

To remedy this situation, Rassoul [18] used extreme value theory taking into account Hill's estimator [19] estimator of the tail index γ and introduced a semi-parametric estimator for the probability of ruin $\phi(u)$ for heavy-tailed losses with infinite second-order moments.

The estimation of γ has been extensively studied in the literature, and γ is a positive-tail index. Hill's estimator is the most popular estimator of the positive-tail index γ in extreme value theory, and is defined as follows:

$$\hat{\gamma}_{1,n,k} := k^{-1} \sum_{j=1}^k j (\log X_{n-j+1,n} - \log X_{n-j,n}),$$

for an intermediate sequence $k = k(n)$, i.e., are sequences such that $k \rightarrow \infty$ and $k/n \rightarrow 0$, as $n \rightarrow \infty$. On the other hand, the Hill estimator associated with the stop loss sample (Y_1, \dots, Y_n) is given as follows:

$$\hat{\gamma}_{2,n,\ell} := \ell^{-1} \sum_{j=1}^{\ell} j (\log Y_{n-j+1,n} - \log Y_{n-j,n}),$$

where $\ell = \ell(n)$ is another intermediate sequences satisfying $\ell \rightarrow \infty$ and $\ell/n \rightarrow 0$, as $n \rightarrow \infty$. In extreme value theory, Hill's estimator has been extensively studied, improved and

even generalized to any real parameter γ (see e.g., [16,25,26]). Its weak consistency was established under the condition of regular variation by Mason [27] assuming only that the underlying distribution varies regularly at infinity. Deheuvels et al. [28] proved the strong consistency of Hill’s estimator.

However, the asymptotic normality of Hill’s estimator has been studied, under various conditions relating to the tail of the distribution, by many researchers, including [25,29,30].

Since the intermediate sequences k and ℓ are such that $k \rightarrow \infty$ and $\ell \rightarrow \infty$, $k/n \rightarrow 0$ and $\ell/n \rightarrow 0$, as $n \rightarrow \infty$; one can decompose respectively the means μ_1 and μ_2 as follows:

$$\mu_1 = \int_0^1 Q_1(s) ds = \int_0^{1-k/n} Q_1(s) ds + \int_{1-k/n}^1 Q_1(s) ds \tag{12}$$

and

$$\mu_2 = \int_0^1 Q_2(s) ds = \int_0^{1-\ell/n} Q_2(s) ds + \int_{1-\ell/n}^1 Q_2(s) ds. \tag{13}$$

Let us define respectively the following estimators for $Q_1(s)$ and $Q_2(s)$, $s \in [0, 1)$:

$$\widehat{Q}_{1,n,k}(s) = \begin{cases} Q_{1,n}(s) & \text{for } 0 \leq s \leq 1 - k/n, \\ Q_{1,n,k}^W(s) & \text{for } 1 - k/n < s < 1, \end{cases}$$

and

$$\widehat{Q}_{2,n,\ell}(s) = \begin{cases} Q_{2,n}(s) & \text{for } 0 \leq s \leq 1 - \ell/n, \\ Q_{2,n,\ell}^W(s) & \text{for } 1 - \ell/n < s < 1, \end{cases}$$

where $Q_{1,n}(\cdot)$ and $Q_{2,n}(\cdot)$ are respectively the empirical quantile estimators of $Q_1(\cdot)$ and $Q_2(\cdot)$ and defined by $Q_{1,n}(s) = X_{j,n}$ and $Q_{2,n}(s) = Y_{j,n}$ for all $s \in ((j - 1)/n, j/n]$, and for all $j = 1, \dots, n$, $Q_{1,n,k}^W(s) = ((1 - s)n/k)^{-\widehat{\gamma}_{1,n,k}^H} X_{n-k,n}$ and $Q_{2,n,\ell}^W(s) = ((1 - s)n/\ell)^{-\widehat{\gamma}_{2,n,\ell}^H} Y_{n-\ell,n}$ are respectively the Weissman’s estimators [31] of high quantiles $Q_1(s)$ and $Q_2(s)$ for $s \rightarrow 1$.

By replacing in (12), $Q_1(s)$ (resp. in (13), $Q_2(s)$) by $\widehat{Q}_{1,n,k}(s)$ (resp. by $\widehat{Q}_{2,n,\ell}(s)$) and by integrating we arrive respectively at the following alternative estimators for the means m_1 and μ_2 when losses are heavy tailed with tail index γ in the upper half of unit interval ($1/2 < \gamma < 1$):

$$\widehat{\mu}_{1,n,k} = n^{-1} \sum_{j=1}^{n-k} X_{j,n} + \frac{k}{n} \frac{X_{n-k,n}}{(1 - \widehat{\gamma}_{1,n,k}^H)},$$

and

$$\widehat{\mu}_{2,n,\ell} = n^{-1} \sum_{j=1}^{n-\ell} Y_{j,n} + \frac{\ell}{n} \frac{Y_{n-\ell,n}}{(1 - \widehat{\gamma}_{2,n,\ell}^H)}.$$

These estimators of means were first studied by Peng [32] and also generalized in [4–6,8] to assess financial and actuarial risk measures.

As in (10), substituting $\widehat{\mu}_{1,n}$ and $\widehat{\mu}_{2,n}$ with μ_1 and μ_2 , respectively, on the right-hand side of the Equation (5), Rassoul [18] introduced the following alternative estimator for the ruin probability $\phi(u)$:

$$\widetilde{\phi}_{n,k,\ell}(u) \sim \frac{\widehat{\mu}_{2,n,\ell}}{\omega - \widehat{\mu}_{1,n,k}} \quad \text{for } 1/2 < \gamma < 1. \quad (14)$$

Rassoul [18] established the asymptotic normality of the estimator $\widetilde{\phi}_{n,k,\ell}(u)$ under certain restrictive assumptions. Finally, an asymptotic normal of $\phi(u)$, for $0 < \gamma < 1$ takes the following form:

$$\widehat{\phi}_n(u) := \begin{cases} \overline{\phi}_n(u), & \text{for } 0 < \gamma \leq 1/2, \\ \widetilde{\phi}_{n,k,\ell}(u), & \text{for } 1/2 < \gamma < 1. \end{cases}$$

Note that the alternative ruin probability estimator $\widetilde{\phi}_{n,k,\ell}(u)$ is associated to the the Hill estimators $\widehat{\gamma}_{1,n,\ell}$ and $\widehat{\gamma}_{2,n,\ell}$.

It is well known that these Hill estimators are both pseudo-maximum likelihood estimators based on the exponential approximation of normalized log-spacings, i.e., $V_{1,j} := j(\log X_{n-j+1,n} - \log X_{n-j,n})$, for $j = 1, \dots, k$ and $V_{2,j} := j(\log Y_{n-j+1,n} - \log Y_{n-j,n})$, for $j = 1, \dots, \ell$, see, e.g., [16,33]. Clearly, these Hill estimators depend respectively on the choice of sample fractions k, ℓ and their influence functions are slowly increasing but not bounded. Consequently, these estimators are not very robust to large values of $V_{\bullet,j}$, which makes the estimator of the probability of ruin $\widetilde{\phi}_{n,k,\ell}(u)$ sensitive. This constitutes a serious problem in terms of bias and mean square error (MSE). To overcome this problem, we introduce in the next section a robust estimator of the probability of ruin $\phi(u)$ for heavy-tailed distributions whose index lies in the upper half of the unit interval, and establish its asymptotic properties.

3. Robust estimator and main results

3.1. Robust estimator for the ruin probability $\phi(u)$

To solve the aforementioned problem of the classical Hill's estimator, Fabián and Stehlík [22] proposed a sensible distribution known as the t-Hill estimator (t-score or score moment estimate) for the tail index of any distribution that varies the tail regularly. In addition, Jordanova and Pancheva [34] discovered the limiting distribution of the t-Hill estimator in the case where the rank $S = k, \ell$ of the higher-order statistic is $o(n)$ and proved its asymptotic normality.

This estimator of the score moment has been studied in [35,36]. According to these authors, this estimator is more robust than the classic Hill estimator. Recently, several studies on t-Hill have been published, see [37,38]. In order to improve the quality of the averages μ_1 and μ_2 given respectively in (12) and (13), which allows us to improve the quality of the infinite-time ruin probability for a heavy-tailed distribution, instead of implementing the Hill's estimator, we propose to estimate the tail index γ by the so-called t-score moment procedure, in order to obtain a robust result.

The formula of the t-Hill estimators of γ are given by:

$$\widehat{\gamma}_{1,n,k}^{tH} =: \left(\frac{1}{k} \sum_{j=1}^k \frac{X_{n-k,n}}{X_{n-j+1,n}} \right)^{-1} - 1 \quad (15)$$

and

$$\widehat{\gamma}_{2,n,\ell}^{tH} =: \left(\frac{1}{\ell} \sum_{j=1}^{\ell} \frac{Y_{n-\ell,n}}{Y_{n-j+1,n}} \right)^{-1} - 1. \quad (16)$$

For other robust estimators of the tail index γ , we may refer the reader to [39–42].

From a Pareto distribution and under mild conditions, Brahim and Kenioua [20] showed in Proposition 1, page 877, that

$$\sqrt{k} (\widehat{\gamma}_{i,n,m}^{tH} - \gamma) \stackrel{d}{=} \gamma (\gamma + 1)^2 \int_0^1 s^{\gamma-1} \mathbb{B}_n(s) ds + o_{\mathbb{P}}(1), \quad i \in \{1, 2\} \text{ and } m \in \{k, \ell\}$$

where $\mathbb{B}_n(s)$, $0 \leq s \leq 1$, is a sequence of Brownian Bridges. This leads to the following asymptotic normal distribution:

$$\sqrt{k} (\widehat{\gamma}_{i,n,m}^{tH} - \gamma) \xrightarrow{d} \mathcal{N}(0, \sigma_{\gamma}^2),$$

where $\sigma_{\gamma}^2 = \gamma^2(\gamma + 1)^2/(2\gamma + 1)$.

As already mentioned, the $\widehat{\phi}_{n,k,\ell}(u)$ estimator given in (5) is not robust. To this end, we provide a solution using the t-Hill estimator of γ to derive a robust estimator of the infinite-time probability of ruin for heavy-tailed $\phi(u)$ distributions. We follow the same method and steps as [18] to write this new estimator, but instead of the simple tail index estimators $\widehat{\gamma}_{1,n,k}^{tH}$ and $\widehat{\gamma}_{2,n,\ell}^{tH}$ defining respectively in (15) and (16), and we introduce the following robust estimators of the ruin probability $\phi(u)$:

$$\widehat{\phi}_{n,k,\ell}^{tH}(u) \sim \frac{\widehat{\mu}_{2,n,\ell}^{tH}}{\omega - \widehat{\mu}_{1,n,k}^{tH}}, \quad \text{for } 1/2 < \gamma < 1, \quad (17)$$

where

$$\widehat{\mu}_{1,n,k}^{tH} := n^{-1} \sum_{j=1}^{n-k} X_{j,n} + \frac{k}{n} \frac{X_{n-k,n}}{(1 - \widehat{\gamma}_{1,n,k}^{tH})}$$

and

$$\widehat{\mu}_{2,n,\ell}^{tH} := n^{-1} \sum_{j=1}^{n-\ell} Y_{j,n} + \frac{\ell}{n} \frac{Y_{n-\ell,n}}{(1 - \widehat{\gamma}_{2,n,\ell}^{tH})}.$$

In the next subsection, we establish the asymptotic properties of our proposed estimator of the probability of ruin.

3.2. Asymptotic results of the estimator $\widehat{\phi}_{n,k,\ell}^{tH}(u)$

As usual in the extreme value framework, to prove asymptotic normality results, we need a second-order condition on the function $\mathbb{U}_i(x) = Q_i(1 - 1/x)$, $x > 1$, $i \in \{1, 2\}$, such as the following:

Condition ($\mathcal{R}_{\mathbb{U}_i}$). There exist a function $A_i(x) \rightarrow 0$ as $x \rightarrow \infty$ of constant sign for large values of x and a second-order parameter $\rho_i < 0$ such that, for every $x > 0$

$$\lim_{t \rightarrow \infty} \frac{\log \mathbb{U}_i(tx) - \log \mathbb{U}_i(t) - \gamma \log(x)}{A_i(t)} = \frac{x^{\rho_i} - 1}{\rho_i}, \quad i \in \{1, 2\}.$$

Note that condition ($\mathcal{R}_{\mathbb{U}_i}$) implies that $|A_i|$ is regularly varying with index ρ_i (see, e.g., [14,43]). It is satisfied for most of the classical distribution functions such as the Pareto, Burr, and Fréchet ones.

Theorem 3.1: Assume that F_i satisfies ($\mathcal{R}_{\mathbb{U}_i}$) with $\gamma \in (1/2, 1)$. Let $k = k(n)$ and $\ell = \ell(n)$ be two intermediate sequences satisfying $k \rightarrow \infty$, $k/n \rightarrow 0$, $\sqrt{k}A_1(n/k) \rightarrow 0$, $\ell \rightarrow \infty$, $\ell/n \rightarrow 0$, $\sqrt{\ell}A_2(n/\ell) \rightarrow 0$ and $\ell/k \rightarrow \theta < \infty$ as $n \rightarrow \infty$. Then on an appropriate probability space, and under a Skorohod construction, there exist a sequence of Brownian bridges $\mathbb{B}_n(s)$, $0 \leq s \leq 1$, such that:

$$\frac{\sqrt{n} \left(\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \right)}{(k/n)^{1/2} X_{n-k,n}} \mathcal{D} = \kappa_1 \sum_{i=1}^3 \mathbb{W}_{n,i} + \theta^{(1/2-\gamma)} \kappa_2 \sum_{i=1}^3 \overline{\mathbb{W}}_{n,i} + o_{\mathbb{P}}(1), \quad (18)$$

where

$$\begin{cases} \mathbb{W}_{n,1} = -\sqrt{\frac{n}{k}} \int_0^{1-\frac{k}{n}} \frac{\mathbb{B}_n(s)}{Q_1\left(1-\frac{k}{n}\right)} dQ_1(s), \\ \mathbb{W}_{n,2} = -\frac{\gamma}{(1-\gamma)} \sqrt{\frac{n}{k}} \mathbb{B}_n\left(1-\frac{k}{n}\right), & \mathbb{W}_{n,3} = \frac{\gamma(\gamma+1)^2}{(1-\gamma)^2} \int_0^1 s^{\gamma-1} \mathbb{B}_n(s) ds, \\ \overline{\mathbb{W}}_{n,1} = -\sqrt{\frac{n}{\ell}} \int_0^{1-\frac{\ell}{n}} \frac{\mathbb{B}_n(s)}{Q_2\left(1-\frac{\ell}{n}\right)} dQ_2(s), \\ \overline{\mathbb{W}}_{n,2} = -\frac{\gamma}{(1-\gamma)} \sqrt{\frac{n}{\ell}} \mathbb{B}_n\left(1-\frac{\ell}{n}\right), \\ \overline{\mathbb{W}}_{n,3} = \frac{\gamma(\gamma+1)^2}{(1-\gamma)^2} \int_0^1 s^{\gamma-1} \mathbb{B}_n(s) ds, \end{cases}$$

with $\kappa_1 = \mu_2/(\omega - \mu_1)^2$, $\kappa_2 = (\omega - \mu_1)^{-1}$ and the construction of the sequence of Brownian Bridges $\mathbb{B}_n(s)$, $0 \leq s \leq 1$ is given in Section 5; statement (21).

Now, by computing the asymptotic variances of the different processes appearing in Theorem 3.1, we deduce the following corollary:

Corollary 3.1: *Under the assumptions of Theorem 3.1, we have*

$$\frac{\sqrt{n} \left(\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \right)}{(k/n)^{1/2} X_{n-k,n}} \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, \sigma^2(\gamma, \theta) \right)$$

where

$$\begin{aligned} \sigma^2(\gamma, \theta) = & \frac{\gamma^2}{(1-\gamma)^2(2\gamma-1)} \left(\kappa_1^2 + \kappa_2^2 \theta^{(1-2\gamma)} + 2\kappa_1\kappa_2 \right) \\ & + \frac{\gamma^2(\gamma+1)^2}{(2\gamma+1)(1-\gamma)^4} \left(\kappa_1 + \kappa_2 \theta^{(1/2-\gamma)} \right)^2. \end{aligned}$$

Remark 3.1: In the case where $k \sim \ell$, as $n \rightarrow \infty$, we have $\theta \sim 1$, one can reduce asymptotic variance $\sigma^2(\gamma, \theta)$ to $4\gamma^5(\kappa_1 + \kappa_2)^2/(1-\gamma)^4(4\gamma^2 - 1)$.

From Corollary 3.1 3.1, the asymptotic variance $\sigma^2(\gamma, \theta)$ of the ruin probability estimator $\widehat{\phi}_{n,k,\ell}^{tH}(u)$ depends on unknown parameters γ and θ . To estimate this asymptotic variance, we recommended to use a block bootstrapping method, which gives the standard deviation and the $(1 - \alpha) \times 100\%$ confidence interval for the estimated value.

By repeating such bootstrapping procedure T times, one can obtain T bootstrapped estimates for the ruin probability estimates. The sample standard deviation across the T estimates gives an estimate of the standard deviation and the of the underlying ruin probability estimators at optimal points of sample fractions $k, \ell \in \{1, \dots, n - 1\}$. The choice of the sample fraction is a serious challenge. Note in this regard that the tail index estimators have large variance when its associated sample fraction is small and large bias when it is large.

To balance between the bias and variance of problem of tail index estimators, several adaptive procedures have been proposed; in our simulations we use that of [9].

One can also estimate the asymptotic variance $\sigma^2(\gamma, \theta)$ by using its plug in estimator $\sigma^2(\widehat{\gamma}, \theta)$, where $\widehat{\gamma}$ is a consistent estimator of the tail index γ . But, this estimation may take on negative values when the tail index γ takes on values smaller than 1/2. This happens quite often when the sample size n is not sufficiently large and the population tail-index γ is only slightly larger than 1/2.

To overcome this problem, We recommend also the following procedure for estimating the asymptotic variance. As we shall show in the proof, the Brownian bridges. As we shall show in the proof, the Brownian bridges in Theorem 1 stem from a Skorokhod-type approximation of the uniform empirical process:

$$\beta_n(t) = \sqrt{n} \left(F_{1,n}(Q_1(t)) - t \right), \quad t \in (0, 1]$$

where $F_{1,n}$ is the empirical component of $F_{1,n}$.

Substituting the Brownian bridges $\mathbb{B}_n(\dots)$ by the uniform empirical process e_n on the right-hand side of Equation (18), we get:

$$\frac{\sqrt{n} \left(\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \right)}{(k/n)^{1/2} X_{n-k,n}} \stackrel{\mathcal{D}}{=} \frac{1}{\sqrt{n}} \sum_{j=1}^n \left(\kappa_1 h_{F_1}(X_j, \gamma) + \theta^{1/2-\gamma} \kappa_2 h_{F_2}(Y_j, \gamma) \right)$$

when $n \rightarrow \infty$, with

$$\begin{aligned} h_{F_1}(X_j, \gamma) &= -\sqrt{\frac{n}{k}} \int_0^{1-\frac{k}{n}} \frac{\mathbb{I}(X_j \geq Q_1(s)) - s}{Q_1\left(1 - \frac{k}{n}\right)} dQ_1(s) \\ &\quad - \frac{\gamma}{(1-\gamma)} \sqrt{\frac{n}{k}} (\mathbb{I}(X_j \geq Q_1(1 - k/n)) - (1 - k/n)) \\ &\quad + \frac{\gamma(\gamma+1)^2}{(1-\gamma)^2} \int_0^1 s^{\gamma-1} (\mathbb{I}(X_j \geq Q_1(s)) - s) ds, \end{aligned}$$

and

$$\begin{aligned} h_{F_2}(Y_j, \gamma) &= -\sqrt{\frac{n}{\ell}} \int_0^{1-\frac{\ell}{n}} \frac{\mathbb{I}(Y_j \geq Q_2(s)) - s}{Q_2\left(1 - \frac{\ell}{n}\right)} dQ_2(s) \\ &\quad - \frac{\gamma}{(1-\gamma)} \sqrt{\frac{n}{\ell}} (\mathbb{I}(Y_j \geq Q_2(1 - \ell/n)) - (1 - \ell/n)) \\ &\quad + \frac{\gamma(\gamma+1)^2}{(1-\gamma)^2} \int_0^1 s^{\gamma-1} (\mathbb{I}(Y_j \geq Q_2(s)) - s) ds, \end{aligned}$$

where the estimated values are given by $h_{F_{1,n}}(X_j, \widehat{\gamma})$ and $h_{F_{2,n}}(Y_j, \widehat{\gamma})$ and are obtained by substituting respectively Q_1 and Q_2 with their empirical components and γ with its consistent estimator as the t-Hill one. So that, the corresponding non-negative estimator of the variance $\sigma^2(\gamma, \theta)$ is given for $\theta \sim \ell/k$ by:

$$\widehat{\sigma}_{n,k,\ell}^2 = \frac{1}{n} \sum_{j=1}^n (\kappa_1 h_{F_{1,n}}(X_j, \widehat{\gamma}) + \theta^{1/2-\gamma} \kappa_2 h_{F_{2,n}}(Y_j, \widehat{\gamma}))^2.$$

4. Simulation study

4.1. Performance and comparative study

In this simulation study, we examine the performance of the new estimator $\widehat{\phi}_{n,k,\ell}^{tH}(u)$ given in (17) with the classical estimator $\widehat{\phi}_{n,k,\ell}(u)$ proposed by Rassoul [18] and defined in (5). Thus, we generate $N = 1000$ samples (X_1, \dots, X_n) with the sample size $n = 1000, 1500, 2000$ from a Pareto distribution function defined as: $F_1(x) = 1 - x^{-1/\gamma}$, $x \geq 1$ with extreme value index $\gamma \in \{2/3, 3/4\}$. For a given initial large reserve $u = 1.5$, we derive for each sample its corresponding excess of loss (Y_1, \dots, Y_n) , where $Y_j = \max(X_j - u, 0)$.

The ruin probability estimators $\widehat{\phi}_{n,k,\ell}^{tH}(u)$ and $\widehat{\phi}_{n,k,\ell}(u)$ are computed with the parameter $\omega = c/\lambda = 18$ (in order to ensure that $\mu_2 < \omega - \mu_1$) and with respectively the tail index estimators $\gamma_{i,n,S}^H$ and $\gamma_{i,n,S}^{tH}$, $(i, S) \in \{(1, k), (2, \ell)\}$, for different sample fractional numbers of top order statistics $k = 1, \dots, n-1$ and $\ell = 1, \dots, m_n-1$, where m_n is the number of positive values of $Y_j \neq 0$, $j = 1, \dots, n$. Employing the algorithm of [9], Page 137, the optimal values k^* and ℓ^* of the number of top extremes of k and ℓ to compute the ruin

probability estimators are respectively values k^* and ℓ^* defined as:

$$k^* = \arg \min_k \frac{1}{k} \sum_{j=1}^k j^\delta \left| \widehat{\gamma}_{1,n,j}^\bullet - \text{median}(\widehat{\gamma}_{1,n,1}^\bullet, \dots, \widehat{\gamma}_{1,n,k}^\bullet) \right|, \quad 1 \leq k \leq n-1, \quad (19)$$

and

$$\ell^* = \arg \min_\ell \frac{1}{\ell} \sum_{j=1}^\ell j^\delta \left| \widehat{\gamma}_{2,n,j}^\bullet - \text{median}(\widehat{\gamma}_{2,n,1}^\bullet, \dots, \widehat{\gamma}_{2,n,\ell}^\bullet) \right|, \quad 1 \leq \ell \leq m_n - 1, \quad (20)$$

where $0 \leq \delta < 1/2$ and $\widehat{\gamma}_{1,n,k}^\bullet$ (respectively $\widehat{\gamma}_{2,n,\ell}^\bullet$, is either the Hill's or the t-Hill's estimator of the tail index γ computed with the sample (X_1, \dots, X_n) , respectively with the excess sample (Y_1, \dots, Y_n) . By the way, choosing $\delta = 1/4$, we compute the optimal values k^* and ℓ^* as in (19) and (20) for each tail index estimator used in the computation of their associated ruin probability estimators.

• Next, we compare the performance of the above-mentioned ruin probability estimators by computing the absolute value of the mean as well as the mean square errors (MSE) based on the $N = 500$ simulated samples, and defined as follows:

$$\text{ABias}(\phi_{n,k^*,\ell^*}^\bullet(u)) := \left| \frac{1}{N} \sum_{j=1}^N \frac{\phi_{n,k^*,\ell^*}^{\bullet,j}(u)}{\phi(u)} - 1 \right|$$

and

$$\text{MSE}(\phi_{n,k^*,\ell^*}^\bullet(u)) := \frac{1}{N} \sum_{j=1}^N \left(\frac{\phi_{n,k^*,\ell^*}^{\bullet,j}(u)}{\phi(u)} - 1 \right)^2,$$

where $\phi(u)$ is the true value of the ruin probability and $\phi_{n,k^*,\ell^*}^{\bullet,j}(u)$ is the j th value ($j = 1, \dots, N$) of any ruin probability estimator $\phi_{n,k^*,\ell^*}^\bullet(u)$ of $\phi(u)$, evaluated at their optimal numbers of higher-order statistics.

The point estimates of the probability of ruin at their optimal k^* values as well as their ABias and MSE are summarized in the following Table 1.

Examination of the table leads to two conclusions, whatever the situation. Firstly, we note that the absolute bias of both probability of ruin estimators decrease to zero when the sample size n becomes large. Secondly, we find that the MSEs of the $\widehat{\phi}_{n,k,\ell}^{tH}(u)$ estimator converge faster to zero as n increases, compared with the MSEs of $\widehat{\phi}_{n,k,\ell}(u)$. In this context, these numerical results show that the estimator $\widehat{\phi}_{n,k,\ell}^{tH}(u)$ is the best.

4.2. Comparative robustness study

One way of increasing robustness is to create a contamination model, which is considered to replace some of the variables of the data X with outliers. Thus, to assess the robustness of our estimator, a simulation was performed with contaminated data for each estimator. The main points here are to consider a ε -contamination model, which consists in considering a Pareto distribution $F_1(x) = 1 - x^{-1/\gamma}$ polluted by variables extracted from the other

Table 1. Estimation results of $\widehat{\phi}_{k^*,\ell^*,n}(u)$ and $\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$ estimators of the ruin probability $\phi(u) = 0.1$ for $u = 1.5$ and $\omega = c/\lambda = 18$, computed with optimal numbers of top statistics k^* and ℓ^* , based on $N = 1000$ samples of size $n = 1000; 1500; 2000$, from the distribution $F_1(x) = 1 - x^{-1/\gamma}$, $\gamma = 2/3; 3/4$.

$\gamma = 2/3, \phi(u) = 0.1$			$\gamma = 3/4, \phi(u) = 0.1576$				
$n = 1000$							
$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.0893	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.0943	$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.1504	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1638
ABais	0.1069	ABais	0.0648	ABais	0.0615	ABais	0.0585
MSE	0.0114	MSE	0.0042	MSE	0.0037	MSE	0.0034
$n = 1500$							
$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.0905	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.0953	$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.1638	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1523
ABais	0.0946	ABais	0.0545	ABais	0.0526	ABais	0.0520
MSE	0.0089	MSE	0.0029	MSE	0.0027	MSE	0.0023
$n = 2000$							
$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.0925	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.0995	$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.1649	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1546
ABais	0.0751	ABais	0.0354	ABais	0.0502	ABais	0.0358
MSE	0.0056	MSE	0.0013	MSE	0.0025	MSE	0.0012

Table 2. Estimation results of $\widehat{\phi}_{k^*,\ell^*,n}(u)$ and $\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$ estimators of the ruin probability ϕ at the point u , computed with optimal numbers of top statistics k^* and ℓ^* , based on $N = 1000$ samples of size $n = 1000; 1500; 2000$, from the distribution $F_1^c(x) = 1 - (1 - \varepsilon)x^{-1/\gamma} + \varepsilon(\frac{x}{a})^{-1/\gamma}$, $\gamma = 2 - 3, 3/4$.

$\gamma = 2/3, \phi(u) = 0.1$			$\gamma = 3/4, \phi(u) = 0.1576008$				
$n = 1000$							
$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.0886	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.0995	$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.1424	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1597
ABais	0.1145	ABais	0.0492	ABais	0.0985	ABais	0.0542
MSE	0.0131	MSE	0.0024	MSE	0.0097	MSE	0.0029
$n = 1500$							
$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.0905	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1012	$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.1724	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1568
ABais	0.0981	ABais	0.0418	ABais	0.0974	ABais	0.0476
MSE	0.0096	MSE	0.0017	MSE	0.00948	MSE	0.0022
$n = 2000$							
$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.0901	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1008	$\widehat{\phi}_{k^*,\ell^*,n}(u)$	0.1422	$\widehat{\phi}_{k^*,\ell^*,n}^{tH}(u)$	0.1567
ABais	0.0947	ABais	0.0305	ABais	0.0939	ABais	0.0348
MSE	0.0089	MSE	0.0012	MSE	0.0088	MSE	0.0013

Pareto distribution $F_{1,a}(x) = 1 - (\frac{x}{a})^{-1/\gamma}$ and to use the following mixing distribution:

$$F_1^c(x) = (1 - \varepsilon) \times F_1(x) + \varepsilon \times F_{1,a}(x) = 1 - (1 - \varepsilon)x^{-1/\gamma} + \varepsilon \left(\frac{x}{a}\right)^{-1/\gamma},$$

where $\varepsilon \in (0, 1)$ is the contamination rate. Now, for a given $\varepsilon = 10\%$ and $a = 3$, we generate also $N = 1000$ samples of size $n = 1000, 1500, 2000$ from the contaminated Pareto distribution $F_1^c(x)$. This kind of ε -contaminated model is used in [20–22] to evaluate insured.

Next, as in Subsection 4.1, we compare the new estimator $\widehat{\phi}_{n,k,\ell}^{tH}(u)$ with the classical estimator $\widehat{\phi}_{n,k,\ell}(u)$ of the ruin probability $\phi(u)$, by calculating the Absolute bias and the MSE. The results are shown in Table 2. It turned out that the effect of contamination gets immediately apparent.

5. Proofs

Let E_1, \dots, E_n be independent and identically distributed random variables from the unit Pareto distribution G , defined as $G(t) = 1 - t^{-1}$, $t \geq 1$. For each $n \geq 1$, let $E_{1,n} \leq \dots \leq E_{n,n}$ be the order statistics pertaining to E_1, \dots, E_n . Clearly $X_{j,n} \stackrel{d}{=} \mathbb{U}_1(E_{j,n})$ and $Y_{j,n} \stackrel{d}{=} \mathbb{U}_2(E_{j,n})$ $j = 1, \dots, n$. In order to be able to applied the results from [44], a probability space $(\Omega, \mathbb{A}, \mathbb{P})$ is constructed carrying a sequence ξ_1, ξ_2, \dots of independent random variables uniformly distributed on $(0, 1)$ and a sequence of Brownian bridges $\mathbb{B}_n(s)$, $0 \leq s \leq 1$, $n = 1, 2, \dots$ such that for all $0 \leq \nu < 1/2$ and $\lambda > 0$

$$\sup_{\lambda/n \leq s \leq 1-\lambda/n} \frac{|\beta_n(s) - \mathbb{B}_n(s)|}{(s(1-s))^{1/2-\nu}} = O_{\mathbb{P}}(n^{-\nu}), \quad (21)$$

where β_n is following the uniform quantile process

$$\beta_n(t) = \sqrt{n}(t - V_n(t))$$

with V_n denoting the empirical uniform quantile function defined to be $V_n(t) = \xi_{j,n}$, $\frac{j-1}{n} < t \leq \frac{j}{n}$, $j = 1, \dots, n$ and $V_n(0) = 0$.

Proof of Theorem 3.1: Recall that:

$$\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \sim \frac{\widehat{\mu}_{2,n,\ell}^{tH}}{\omega - \widehat{\mu}_{1,n,k}^{tH}} - \frac{\mu_2}{\omega - \mu_1},$$

where

$$\widehat{\mu}_{1,n,k}^{tH} := \int_0^{1-k/n} Q_{1,n}(s) ds + \frac{k}{n} \frac{X_{n-k,n}}{1 - \widehat{\gamma}_{1,n,k}^{tH}},$$

and

$$\widehat{\mu}_{2,n,\ell}^{tH} := \int_0^{1-\ell/n} Q_{2,n}(s) ds + \frac{\ell}{n} \frac{Y_{n-\ell,n}}{1 - \widehat{\gamma}_{2,n,\ell}^{tH}}.$$

Then, one may be rewrite the random variable $\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u)$ as follows:

$$\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \sim \frac{\mu_2}{(\omega - \widehat{\mu}_{1,n,k}^{tH})(\omega - \mu_1)} (\widehat{\mu}_{1,n,k}^{tH} - \mu_1) + \frac{1}{\omega - \widehat{\mu}_{1,n,k}^{tH}} (\widehat{\mu}_{2,n,\ell}^{tH} - \mu_2).$$

Under the assumption (7), the application of Theorem 2.4.1 in [14], ensures that, $X_{n-k,n} = Q_1(1 - k/n)\{1 + o_{\mathbb{P}}(1)\}$, as $n \rightarrow \infty$. Since the relation (7) is equivalent to $Q_1(1 - s) = s^{-\gamma} \ell_{Q_1}(s)$, $s \in (0, 1)$, where ℓ_{Q_1} is a slowly varying function, more precisely $\ell_{Q_1}(sx)/\ell_{Q_1}(s) \rightarrow 1$, as $s \rightarrow 0$, then $(k/n)Q_1(1 - k/n) = (k/n)^{1-\gamma} \ell_{Q_1}(k/n)$. Thus, for a given $\gamma \in (1/2, 1)$, we have from Proposition 1.3.6 in [11], $(k/n)^{1-\gamma} \ell_{Q_1}(k/n) \rightarrow 0$, as $n \rightarrow \infty$. Therefore, using the weak consistency of the estimator $\widehat{\gamma}_{1,n,k}^{tH}$ to γ (see, Jordanova et al. [38]), we obtain $(k/n)X_{n-k,n}/(1 - \widehat{\gamma}_{1,n,k}^{tH}) \xrightarrow{\mathbb{P}} 0$, as $n \rightarrow \infty$.

Also, under assumption we have from [32],

$$\frac{\sqrt{n} \int_0^{1-k/n} (Q_{1,n}(s) - Q_1(s)) ds}{(k/n)^{1/2} X_{n-k,n}} \stackrel{d}{=} -\frac{\int_0^{1-k/n} \mathbb{B}_n(s) dQ_1(s)}{(k/n)^{1/2} Q_1(1-k/n)} + o_{\mathbb{P}}(1). \quad (22)$$

Since the right term in (22) is bounded in probability, it comes for all large values of n ,

$$\int_0^{1-k/n} Q_{1,n}(s) ds = \int_0^{1-k/n} Q_1(s) ds + o_{\mathbb{P}}(1).$$

Remarking that $k/n \rightarrow 0$, as $n \rightarrow \infty$, we have $\int_0^{1-k/n} Q_1(s) ds = \int_0^1 Q_1(s) ds \{1 + o_{\mathbb{P}}(1)\}$, as $n \rightarrow \infty$. Which leads to the convergence in probability of $\widehat{\mu}_{1,n,k}^{tH}$ to the mean μ_1 . Next, let's denote by $\kappa_1 = \mu_2/(\omega - \mu_1)^2$ and $\kappa_2 = 1/(\omega - \mu_1)$. Then, for all n large enough, the random variable $\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u)$ can be also represented as follows

$$\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \stackrel{d}{=} \kappa_1 (\widehat{\mu}_{1,n,k}^{tH} - \mu_1) + \kappa_2 (\widehat{\mu}_{2,n,\ell}^{tH} - \mu_2).$$

Consequently,

$$\begin{aligned} \frac{\sqrt{n} \left(\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \right)}{(k/n)^{1/2} X_{n-k,n}} &\stackrel{d}{=} \kappa_1 \frac{n^{1/2} \left(\widehat{\mu}_{1,n,k}^{tH} - \mu_1 \right)}{(k/n)^{1/2} Q_1(1-k/n)} \\ &\quad + \kappa_2 \sqrt{\ell/k} \left(\frac{Q_2(1-\ell/n)}{Q_1(1-k/n)} \right) \frac{n^{1/2} \left(\widehat{\mu}_{2,n,\ell}^{tH} - \mu_2 \right)}{(\ell/n)^{1/2} Q_2(1-\ell/n)}, \\ &:= A_{1,n} + A_{2,n}. \end{aligned}$$

Now, let We first compute the $A_{1,n}$ term. By substituting $\widehat{\mu}_{1,n,k}^{tH}$ and μ_1 with their expressions, we have

$$\begin{aligned} A_{1,n} &= \kappa_1 \frac{\sqrt{n} \int_0^{1-k/n} (Q_{1,n}(s) - Q_1(s)) ds}{(k/n)^{1/2} Q_1(1-k/n)} \\ &\quad + \kappa_1 \frac{n^{1/2} \left(\frac{(k/n)X_{n-k,n}}{1-\widehat{\gamma}_{1,n,k}^{tH}} - \int_{1-k/n}^1 Q_1(s) ds \right)}{(k/n)^{1/2} Q_1(1-k/n)} \\ &:= A_{1,n}^{(1)} + A_{1,n}^{(2)}. \end{aligned}$$

From the statement in (22), we have

$$A_{1,n}^{(1)} \stackrel{d}{=} -\kappa_1 \mathbb{W}_{n,1} + o_{\mathbb{P}}(1), \quad (23)$$

where

$$\mathbb{W}_{n,1} := \frac{\int_0^{1-k/n} \mathbb{B}_n(s) dQ_1(s)}{(k/n)^{1/2} Q_1(1-k/n)}.$$

Next, remarking that $\mathbb{U}_1(n/k) = Q_1(1-k/n)$ and $X_{n-k,n} \stackrel{d}{=} \mathbb{U}_1(E_{n-k,n})$, we have

$$A_{1,n}^{(2)} \stackrel{d}{=} \sum_{i=1}^4 T_{n,i},$$

where

$$\begin{aligned} T_{n,1} &= \frac{\kappa_1}{1 - \widehat{\gamma}_{1,n,k}^{tH}} \sqrt{k} \left[\frac{\mathbb{U}_1(E_{n-k,n})}{\mathbb{U}_1(n/k)} - \left(\frac{k}{n} E_{n-k,n} \right)^\gamma \right], \\ T_{n,2} &= \frac{\kappa_1}{1 - \widehat{\gamma}_{1,n,k}^{tH}} \sqrt{k} \left(\left(\frac{k}{n} E_{n-k,n} \right)^\gamma - 1 \right), \\ T_{n,3} &= \frac{\kappa_1}{(1 - \gamma) (1 - \widehat{\gamma}_{1,n,k}^{tH})} \sqrt{k} (\widehat{\gamma}_{1,n,k}^{tH} - \gamma), \\ T_{n,4} &= \kappa_1 \sqrt{k} \left[\frac{1}{1 - \gamma} - \frac{\int_1^{+\infty} s^{-2} \mathbb{U}_1(ns/k) ds}{\mathbb{U}_1(n/k)} \right]. \end{aligned}$$

We study each term separately.

Term $T_{n,1}$. According to [14], Theorem 2.3.9), for any $\delta > 0$, we have

$$\begin{aligned} & \sqrt{k} \left(\frac{\mathbb{U}_1(E_{n-k,n})}{\mathbb{U}_1(n/k)} - \left(\frac{k}{n} E_{n-k,n} \right)^\gamma \right) \\ &= \sqrt{k} A_1 \left(\frac{n}{k} \right) \left\{ \left(\frac{k}{n} E_{n-k,n} \right)^\gamma \frac{\left(\frac{k}{n} E_{n-k,n} \right)^\rho - 1}{\rho} + o_{\mathbb{P}}(1) \left(\frac{k}{n} E_{n-k,n} \right)^{\gamma + \rho \pm \delta} \right\}. \end{aligned}$$

Thus, since $kE_{n-k,n}/n \rightarrow 1$, $\sqrt{k} A_1(n/k) \rightarrow 0$ and $\widehat{\gamma}_{1,n,k}^{tH} \xrightarrow{\mathbb{P}} \gamma$, as $n \rightarrow \infty$ (see, Jordanova et al. [38]), it readily follows that

$$T_{n,1} = o_{\mathbb{P}}(1). \quad (24)$$

Term $T_{n,2}$. The equality $E_{n-k,n} \stackrel{d}{=} (1 - \xi_{n-k,n})^{-1}$ yields:

$$\begin{aligned} & \sqrt{k} \left[\left(\frac{k}{n} E_{n-k,n} \right)^\gamma - 1 \right] \\ & \stackrel{d}{=} \sqrt{k} \left(\left(\frac{n}{k} (1 - \xi_{n-k,n}) \right)^{-\gamma} - 1 \right) \\ & = -\gamma \sqrt{k} \left(\frac{n}{k} (1 - \xi_{n-k,n}) - 1 \right) (1 + o_{\mathbb{P}}(1)) \quad \text{by a Taylor expansion} \\ & = -\gamma \sqrt{\frac{n}{k}} \beta_n \left(1 - \frac{k}{n} \right) (1 + o_{\mathbb{P}}(1)) \\ & = -\gamma \sqrt{\frac{n}{k}} \left(\mathbb{B}_n \left(1 - \frac{k}{n} \right) + O_{\mathbb{P}}(n^{-\nu}) \left(\frac{k}{n} \right)^{1/2-\nu} \right) (1 + o_{\mathbb{P}}(1)), \end{aligned}$$

for $0 \leq \nu < 1/2$, by Csörgő et al. [44]. Thus, using again the weak consistency of $\widehat{\gamma}_{1,n,k}^{tH}$ to γ , it follows that:

$$T_{n,2} \stackrel{d}{=} \kappa_1 \mathbb{W}_{n,2} (1 + o_{\mathbb{P}}(1)), \tag{25}$$

where

$$\mathbb{W}_{n,2} = -\frac{\gamma}{(1-\gamma)} \sqrt{\frac{n}{k}} \mathbb{B}_n \left(1 - \frac{k}{n}\right).$$

Term $T_{n,3}$. From the Proposition 1 in [20], page 877, we have

$$\sqrt{k} (\widehat{\gamma}_{1,n,k}^{tH} - \gamma) \stackrel{d}{=} \gamma (\gamma + 1)^2 \int_0^1 s^{\gamma-1} \mathbb{B}_n(s) ds + o_{\mathbb{P}}(1).$$

And by using again the consistency in probability of $\widehat{\gamma}_{1,n,k}^{tH}$ to γ , we get for all n large enough:

$$T_{n,3} = \kappa_1 \frac{\gamma (\gamma + 1)^2}{(1-\gamma)^2} \int_0^1 s^{\gamma-1} \mathbb{B}_n(s) ds + o_{\mathbb{P}}(1) = \kappa_1 \mathbb{W}_{n,3} + o_{\mathbb{P}}(1). \tag{26}$$

Term $T_{n,4}$. A change of variables and an integration by parts yield

$$\begin{aligned} T_{n,4} &= \kappa_1 \sqrt{k} \left\{ \frac{1}{1-\gamma} - \int_1^\infty x^{-2} \frac{\mathbb{U}_1(nx/k)}{\mathbb{U}_1(n/k)} dx \right\} \\ &= -\kappa_1 \sqrt{k} \int_1^\infty x^{-2} \left(\frac{\mathbb{U}_1(nx/k)}{\mathbb{U}_1(n/k)} - x^\gamma \right) dx. \end{aligned}$$

Theorem 2.3.9 in [14] entails that, for $\gamma \in (1/2, 1)$,

$$\begin{aligned} T_{n,4} &= -\kappa_1 \sqrt{k} A_1 \left(\frac{n}{k}\right) \int_1^\infty x^{\gamma-2} \frac{x^\rho - 1}{\rho} dx (1 + o_{\mathbb{P}}(1)) \\ &= \kappa_1 \sqrt{k} A_1 \left(\frac{n}{k}\right) \frac{1}{(1-\gamma)(\gamma + \rho - 1)} (1 + o_{\mathbb{P}}(1)), \\ &= o_{\mathbb{P}}(1), \quad \text{by } \sqrt{k} A_1(n/k) \rightarrow 0. \end{aligned} \tag{27}$$

Combining (23), (24), (25), (26) and (27), we get

$$A_{n,1} \stackrel{d}{=} \kappa_1 (\mathbb{W}_{n,1} + \mathbb{W}_{n,2} + \mathbb{W}_{n,3}) + o_{\mathbb{P}}(1). \tag{28}$$

Notice that for fixed u , we have respectively, from (3) and (7), $1 - F_1(x) = x^{-1/\gamma} (1 + o(1))$ and $1 - F_2(x) = x^{-1/\gamma} (1 + o(1))$, as $x \rightarrow \infty$. This leads to

$$1 - F_1(x) \sim 1 - F_2(x) \quad \text{as } x \rightarrow \infty,$$

and therefore

$$Q_1(1-s) \sim Q_2(1-s) \quad \text{as } s \downarrow 0.$$

Now let's compute the $A_{2,n}$ term. We first have: $Q_2(1 - \frac{\ell}{n}) \sim Q_1(1 - \frac{\ell}{n}) = Q_1(1 - \frac{\ell k}{n}) \sim (\frac{\ell}{k})^{-\gamma} Q_1(1 - \frac{k}{n})$, as $n \rightarrow \infty$. Since by assumption $\ell/k \rightarrow \theta > 0$, then $Q_2(1 - \frac{\ell}{n}) \sim$

$\theta^{-\gamma} Q_1(1 - \frac{k}{n})$, as $n \rightarrow \infty$. This leads for large values of n to

$$A_{2,n} \stackrel{d}{=} \theta^{(1/2-\gamma)} \kappa_2 \frac{n^{1/2} (\widehat{\mu}_{2,n,\ell}^{tH} - \mu_2)}{(\ell/n)^{1/2} Q_2(1 - \ell/n)}.$$

Further, Substituting $\widehat{\mu}_{2,n,\ell}^{tH}$ and μ_2 with their expressions and using similar arguments as those developed to show the expression $A_{1,n}$ in (28) together with $Y_{n-\ell,n} \stackrel{d}{=} \mathbb{U}_2(E_{n-\ell,n})$, it comes:

$$A_{n,2} \stackrel{d}{=} \theta^{(1/2-\gamma)} \kappa_2 \left(\overline{\mathbb{W}}_{n,1} + \overline{\mathbb{W}}_{n,2} + \overline{\mathbb{W}}_{n,3} \right) + o_{\mathbb{P}}(1), \quad (29)$$

where

$$\begin{aligned} \overline{\mathbb{W}}_{n,1} &:= -\frac{\int_0^{1-\ell/n} \mathbb{B}_n(s) dQ_2(s)}{(\ell/n)^{1/2} Q_2(1 - \ell/n)}, \\ \overline{\mathbb{W}}_{n,2} &:= -\frac{\gamma}{1-\gamma} \sqrt{\frac{n}{\ell}} \mathbb{B}_n \left(1 - \frac{\ell}{n} \right), \\ \overline{\mathbb{W}}_{n,3} &:= \frac{\gamma(\gamma+1)^2}{(1-\gamma)^2} \int_0^1 s^{\gamma-1} \mathbb{B}_n(s) ds. \end{aligned}$$

Finally, for all n large enough, the Theorem 3.1 holds with

$$\frac{\sqrt{n} \left(\widehat{\phi}_{n,k,\ell}^{tH}(u) - \phi(u) \right)}{(k/n)^{1/2} X_{n-k,n}} \stackrel{d}{=} \kappa_1 \sum_{i=1}^3 \mathbb{W}_{n,i} + \theta^{(1/2-\gamma)} \kappa_2 \sum_{i=1}^3 \overline{\mathbb{W}}_{n,i} + o_{\mathbb{P}}(1).$$

■

Proof of Corollary 3.1: In this part, we use the same approach as in [5]. Since $\mathbb{W}_{n,i}$ and $\overline{\mathbb{W}}_{n,i}$, $i = 1, 2, 3$ are centred Gaussian process, then from Theorem 3.1, we just need to compute the asymptotic variance $\sigma^2(\gamma, \theta)$ of the limiting process. More precisely, have

$$\begin{aligned} \sigma^2(\gamma, \theta) &= \lim_{n \rightarrow \infty} \left((\kappa_1)^2 \left\{ \mathbb{E}(\mathbb{W}_{n,1}^2) + \mathbb{E}(\mathbb{W}_{n,2}^2) + \mathbb{E}(\mathbb{W}_{n,3}^2) \right\} \right. \\ &\quad + (\kappa_2)^2 \theta^{(1-2\gamma)} \left\{ \mathbb{E}(\overline{\mathbb{W}}_{n,1}^2) + \mathbb{E}(\overline{\mathbb{W}}_{n,2}^2) + \mathbb{E}(\overline{\mathbb{W}}_{n,3}^2) \right\} \\ &\quad + 2(\kappa_1)^2 \left\{ \mathbb{E}(\mathbb{W}_{n,1} \mathbb{W}_{n,2}) + \mathbb{E}(\mathbb{W}_{n,1} \mathbb{W}_{n,3}) + \mathbb{E}(\mathbb{W}_{n,2} \mathbb{W}_{n,3}) \right\} \\ &\quad + 2(\kappa_2)^2 \theta^{(1-2\gamma)} \left\{ \mathbb{E}(\overline{\mathbb{W}}_{n,1} \overline{\mathbb{W}}_{n,2}) + \mathbb{E}(\overline{\mathbb{W}}_{n,1} \overline{\mathbb{W}}_{n,3}) + \mathbb{E}(\overline{\mathbb{W}}_{n,2} \overline{\mathbb{W}}_{n,3}) \right\} \\ &\quad + 2\kappa_1 \kappa_2 \theta^{(1/2-\gamma)} \left\{ \mathbb{E}(\mathbb{W}_{n,1} \overline{\mathbb{W}}_{n,1}) + \mathbb{E}(\mathbb{W}_{n,1} \overline{\mathbb{W}}_{n,2}) + \mathbb{E}(\mathbb{W}_{n,1} \overline{\mathbb{W}}_{n,3}) \right\} \\ &\quad + 2\kappa_1 \kappa_2 \theta^{(1/2-\gamma)} \left\{ \mathbb{E}(\mathbb{W}_{n,2} \overline{\mathbb{W}}_{n,1}) + \mathbb{E}(\mathbb{W}_{n,2} \overline{\mathbb{W}}_{n,2}) + \mathbb{E}(\mathbb{W}_{n,2} \overline{\mathbb{W}}_{n,3}) \right\} \\ &\quad \left. + 2\kappa_1 \kappa_2 \theta^{(1/2-\gamma)} \left\{ \mathbb{E}(\mathbb{W}_{n,3} \overline{\mathbb{W}}_{n,1}) + \mathbb{E}(\mathbb{W}_{n,3} \overline{\mathbb{W}}_{n,2}) + \mathbb{E}(\mathbb{W}_{n,3} \overline{\mathbb{W}}_{n,3}) \right\} \right). \end{aligned}$$

Using the Lemma 6 in [4] and following the proof of Corollary 1 (p. 118–119) in [5] with the fact that $\ell/k \rightarrow \infty$, as $n \rightarrow \infty$, $Q_1(1 - su)/Q_1(1 - u) \rightarrow s^{-\gamma}$, $Q_2(1 - su)/Q_2(1 - u) \rightarrow$

$s^{-\gamma}$ and $Q_1(1-u) \sim Q_2(1-u)$ as $u \rightarrow 0$, we get, as $n \rightarrow \infty$:

$$\begin{aligned} \mathbb{E}(\mathbb{W}_{n,1}^2) &\longrightarrow \frac{2\gamma}{2\gamma-1}, & \mathbb{E}(\mathbb{W}_{n,2}^2) &\longrightarrow \frac{\gamma^2}{(1-\gamma)^2}, \\ \mathbb{E}(\mathbb{W}_{n,3}^2) &= \frac{\gamma^2(\gamma+1)^2}{(2\gamma+1)(\gamma-1)^4}, & \mathbb{E}(\mathbb{W}_{n,1}\mathbb{W}_{n,2}) &\longrightarrow \frac{\gamma}{1-\gamma}, \\ \mathbb{E}(\mathbb{W}_{n,1}\mathbb{W}_{n,3}) &= 0 + o(1), & \mathbb{E}(\mathbb{W}_{n,2}\mathbb{W}_{n,3}) &= 0 + o(1), \\ \mathbb{E}(\overline{\mathbb{W}}_{n,1}^2) &\longrightarrow \frac{2\gamma}{2\gamma-1}, & \mathbb{E}(\overline{\mathbb{W}}_{n,2}^2) &\longrightarrow \frac{\gamma^2}{(1-\gamma)^2}, \\ \mathbb{E}(\overline{\mathbb{W}}_{n,3}^2) &= \frac{\gamma^2(\gamma+1)^2}{(2\gamma+1)(\gamma-1)^4}, & \mathbb{E}(\overline{\mathbb{W}}_{n,1}\overline{\mathbb{W}}_{n,2}) &\longrightarrow \frac{\gamma}{1-\gamma}, \\ \mathbb{E}(\overline{\mathbb{W}}_{n,1}\overline{\mathbb{W}}_{n,3}) &= 0 + o(1), & \mathbb{E}(\overline{\mathbb{W}}_{n,2}\overline{\mathbb{W}}_{n,3}) &= 0 + o(1), \\ \mathbb{E}(\mathbb{W}_{n,1}\overline{\mathbb{W}}_{n,1}) &\longrightarrow \frac{\gamma}{1-\gamma} \left(\frac{1}{2\gamma-1} - \theta^{(1-\gamma)} \right) \theta^{(\gamma-1/2)}, \\ \mathbb{E}(\mathbb{W}_{n,1}\overline{\mathbb{W}}_{n,2}) &\longrightarrow \frac{\gamma}{1-\gamma} \theta^{1/2}, \\ \mathbb{E}(\mathbb{W}_{n,1}\overline{\mathbb{W}}_{n,3}) &= 0 + o(1), & \mathbb{E}(\mathbb{W}_{n,2}\overline{\mathbb{W}}_{n,1}) &\longrightarrow \frac{\gamma(1-\gamma\theta^{(1-\gamma)})}{(1-\gamma)^2} \theta^{(\gamma-1/2)}, \\ \mathbb{E}(\mathbb{W}_{n,2}\overline{\mathbb{W}}_{n,2}) &\longrightarrow \frac{\gamma^2}{(1-\gamma)^2} \theta^{1/2}, & \mathbb{E}(\mathbb{W}_{n,2}\overline{\mathbb{W}}_{n,3}) &= 0 + o(1), \\ \mathbb{E}(\mathbb{W}_{n,3}\overline{\mathbb{W}}_{n,1}) &= 0 + o(1), & \mathbb{E}(\mathbb{W}_{n,3}\overline{\mathbb{W}}_{n,2}) &= 0 + o(1), & \mathbb{E}(\mathbb{W}_{n,3}\overline{\mathbb{W}}_{n,3}) \\ &= \frac{\gamma^2(1+\gamma)^2}{(1+2\gamma)(1-\gamma)^4}. \end{aligned}$$

■

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- [1] Panjer H, Willmot GE. Insurance risk models. Michigan: Society of actuaries; 1992.
- [2] Čížek P, Härdle W, Weron R. Statistical tools for finance and insurance. Berlin: Springer; 2005.
- [3] Vandewalle B, Beirlant J. On univariate extreme value statistics and the estimation of reinsurance premiums. Insurance Math Econom. 2006;38:441–459. doi: [10.1016/j.insmatheco.2005.11.002](https://doi.org/10.1016/j.insmatheco.2005.11.002)
- [4] Deme E, Girard S, Guillou A. Reduced-bias estimator of the proportional hazard premium for heavy-tailed distributions. Insurance Math Econom. 2013;52:550–559. doi: [10.1016/j.insmatheco.2013.03.010](https://doi.org/10.1016/j.insmatheco.2013.03.010)
- [5] Deme EH, Girard S, Guillou A. Reduced-biased estimators of the conditional tail expectation for heavy-tailed distributions. Springer; 2015. (Mathematical statistics and limit theorems; p. 105–123).

- [6] Deme EH, Allaya M, Deme S, et al. Estimation of risk measures from heavy-tailed distributions. *Far East J Theor Stat.* 2021;62(1):35–80. doi: [10.17654/TS062010035](https://doi.org/10.17654/TS062010035)
- [7] Matthys G, Delafosse E, Guillou A, et al. Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance Math Econom.* 2004;34:517–537. doi: [10.1016/j.insmatheco.2004.03.004](https://doi.org/10.1016/j.insmatheco.2004.03.004)
- [8] Necir A, Rassoul A, Zitikis R. Estimating the conditional tail expectation in the case of heavy-tailed losses. *J Probab Stat.* 2010;2010(1):Article ID 596839, 17 pp.
- [9] Reiss RD, Thomas M. *Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields.* 3rd ed. Basel, Boston, Berlin: Birkhäuser; 2007.
- [10] Beirlant J, Matthys G, Dierckx G. Heavy-tailed distributions and rating. *Astin Bull.* 2001;31:37–58. doi: [10.2143/AST.31.1.993](https://doi.org/10.2143/AST.31.1.993)
- [11] Bingham NH, Goldie CM, Teugels JL. *Regular variation.* Cambridge: Cambridge University Press; 1989.
- [12] Rolski T, Schmidli H, Schimd V, et al. *Stochastic processes for insurance and finance.* New York: John Wiley and Sons; 1999.
- [13] Asmussen S. *Ruin probabilities.* Singapore: World Scientific; 2000.
- [14] de Haan L, Ferreira A. *Extreme value theory: an introduction.* Springer Series in Operations Research and Financial Engineering; 2006.
- [15] Jones BL, Zitikis R. Risk measures, distortion parameters and their empirical estimation. *Insurance Math Econom.* 2007;41(2):279–297. doi: [10.1016/j.insmatheco.2006.11.001](https://doi.org/10.1016/j.insmatheco.2006.11.001)
- [16] Beirlant J, Dierckx GY, Matthys G. Tail index estimation and an exponential regression model. *Extremes.* 1999;2:177–200. doi: [10.1023/A:1009975020370](https://doi.org/10.1023/A:1009975020370)
- [17] Beirlant J, Dierckx D, Guillou A, et al. On exponential representations of log-spacings of extreme order statistics. *Extremes.* 2002;5(2):157–180. doi: [10.1023/A:1022171205129](https://doi.org/10.1023/A:1022171205129)
- [18] Rassoul A. Estimation of the ruin probability in infinite time for heavy right-tailed losses; 2014.
- [19] Hill BM. A simple approach to inference about the tail of a distribution. *Ann Stat.* 1975;3:1136–1174. doi: [10.1214/aos/1176343247](https://doi.org/10.1214/aos/1176343247)
- [20] Brahim B, Kenioua Z. Robust estimator of distortion risk premiums for heavy-tailed losses. *Afrika Statistika.* 2016;11(1):869–882. doi: [10.16929/as](https://doi.org/10.16929/as)
- [21] Bouali DL, Bouali B, Chesneau C. Robust estimator of conditional tail expectation of Pareto-type distribution. *J Stat Theory Pract.* 2021;15:16. doi: [10.1007/s42519-020-00153-0](https://doi.org/10.1007/s42519-020-00153-0)
- [22] Fabián Z, Stehlík M. On robust and distribution sensitive Hill like method; 2009. (IFAS research report; 43).
- [23] Jordanova PK, Fabián Z, Štěřelec L. On estimation and testing for Pareto tails. *Pliska Stud Math Bulgar.* 2013;22(1):89–108.
- [24] Shorack G, Wellner J. *Empirical processes with applications to statistics.* New York: John Wiley Sons; 1986.
- [25] Dekkers A, Einmahl J, de Haan L. A moment estimator for the index of an extreme-value distribution. *Ann Stat.* 1989;17:1833–1855.
- [26] Lo GS, Fall AM. Another look at second order condition in extreme value theory. *Afr Stat.* 2011;6:346–370.
- [27] Mason D. Laws of large numbers for sums of extreme values. *Ann Probab.* 1982;10:754–764.
- [28] Deheuvels P, Haeusler E, Mason D. Almost sure convergence of the Hill estimator. *Math Proc Camb Philos Soc.* 1988;104:371–381. doi: [10.1017/S0305004100065531](https://doi.org/10.1017/S0305004100065531)
- [29] Beirlant J, Teugels J. Asymptotic normality of Hill’s estimator. In: *Extreme value theory (Oberwolfach, 1987); Lecture notes in statist. Vol. 51;* 1989. p. 148–155.
- [30] Csörgő S, Deheuvels P, Mason DM. Kernel estimates of the tail index of a distribution. *Ann Stat.* 1985;13:1050–1077.
- [31] Weissman I. Estimation of parameters and large quantiles based on the k largest observations. *J Amer Stat Assoc.* 1978;73:812–815.
- [32] Peng L. Estimating the mean of a heavy tailed distribution. *Stat Probab Lett.* 2001;52:255–264. doi: [10.1016/S0167-7152\(00\)00203-0](https://doi.org/10.1016/S0167-7152(00)00203-0)
- [33] Beirlant J, Goegebeur Y, Teugels J, et al. *Statistics of extremes: theory and applications.* John Wiley and Sons Ltd; 2004. (Wiley series in probability and statistics).

- [34] Jordanova PK, Pancheva EI. Weak asymptotic results for t-Hill estimator. *C R Acad Bulgare Sci.* 2012;65(12):1649–1656.
- [35] Stehlík M, Potocký R, Waldl H, et al. On the favorable estimation for fitting heavy tailed data. *Comput Stat.* 2010;25(3):485–503. doi: [10.1007/s00180-010-0189-1](https://doi.org/10.1007/s00180-010-0189-1)
- [36] Stehlík M, Fabián Z, Střelec L. Small sample robust testing for normality against Pareto tails. *Comm Stat Simulat Comput.* 2012;41(7):1167–1194. doi: [10.1080/03610918.2012.625849](https://doi.org/10.1080/03610918.2012.625849)
- [37] Beran J, Schell D, Stehlík M. The harmonic moment tail index estimator: asymptotic distribution and robustness. *Ann Inst Stat Math* 2014;66(1):193–220. doi: [10.1007/s10463-013-0412-2](https://doi.org/10.1007/s10463-013-0412-2)
- [38] Jordanova P, Fabián Z, Hermann P, et al. Weak properties and robustness of t-Hill estimators. *Extremes.* 2016;19(4):591–626. doi: [10.1007/s10687-016-0256-2](https://doi.org/10.1007/s10687-016-0256-2)
- [39] Juárez SF, Schucany WR. Robust and efficient estimation for the generalized Pareto distribution. *Extremes.* 2004;7(3):237–251. doi: [10.1007/s10687-005-6475-6](https://doi.org/10.1007/s10687-005-6475-6)
- [40] Kim M, Lee S. Estimation of a tail index based on minimum density power divergence. *J Multivariate Anal.* 2008;99(10):2453–2471. doi: [10.1016/j.jmva.2008.02.031](https://doi.org/10.1016/j.jmva.2008.02.031)
- [41] Peng L, Welsh AH. Robust estimation of the generalized Pareto distribution. *Extremes.* 2001;4(1):53–65. doi: [10.1023/A:1012233423407](https://doi.org/10.1023/A:1012233423407)
- [42] Vandewalle B, Beirlant J, Christmann A, et al. A robust estimator for the tail index of Pareto-type distributions. *Comput Stat Data Anal.* 2007;51(12):6252–6268. doi: [10.1016/j.csda.2007.01.003](https://doi.org/10.1016/j.csda.2007.01.003)
- [43] Geluk JL, de Haan L. Regular variation, extensions and Tauberian theorems: CWI tract 40. Amsterdam (AB), The Netherlands: Center for Mathematics and Computer Science; 1987.
- [44] Csörgő M, Csörgő S, Horváth L, et al. Weighted empirical and quantile processes. *Ann Probab.* 1986;14:31–85.



Parameter Estimation for Some Discretely Observed Class of Stable Driven Stochastic Differential Equations

Solym M. Manou-Abi^{1,2,3}

Accepted: 2 April 2024

© The Indian Society for Probability and Statistics (ISPS) 2024

Abstract

In this paper, we consider the problem of parameter estimation for a real stochastic model observed at some discrete times, that is a solution of a stochastic differential equation driven by α -stable processes, $\alpha \in (1, 2)$. After recalling the non-parametric estimation framework of the drift function namely the Nadaraya-Watson estimation, we provide explicit estimators for the diffusion parameters (the scaling and the driving stable process parameters) based on the Euler-Maruyama scheme. We apply the estimation results to stable driven Ornstein-Uhlenbeck (OU), Cox-Ingersoll-Ross (CIR) and Lotka-Volterra processes. We also consider the estimation of the drift coefficients in the linear case namely, the stable driven OU and CIR processes. The novelty of this paper which is our baseline is the combination of a characteristic sample function method, the least squares or linear statistical regression methods and the Itô formula. We also established under certain conditions, the consistency of their drift coefficient estimators of the stable driven OU and CIR processes. We efficiently discuss our result with numerical simulations using synthetic data. A real data in finance, such as exchange rates is used to fit the parameters of a justified model among the above stable driven processes. As a forthcoming work, we intend to study the rate of convergence of the estimators and to create a package on R software to handle this kind of estimation problem. We are also currently interested in ergodicity properties for a class of stochastic differential equations driven by stable processes.

Keywords Nadaraya-Watson estimation · Stable process · Stochastic differential equation · Characteristic function · Regression method · CIR · OU and Lotka-Volterra processes

Extended author information available on the last page of the article

1 Introduction

Due to an increase of powerful computation of statistical methods, there has been a great interest in parameter estimation for Stochastic Differential Equation (SDE). Such models which are mathematical tools for modelling the time evolution of several natural phenomena in many fields like epidemiology, biology and finance. For instance, the Cox-Ingersoll-Ross (CIR) and the Ornstein-Uhlenbeck (OU) stochastic models (Cox et al. 2005; Maller et al. 2009), have been broadly used in finance. Furthermore, it is worth noting that select stochastic differential equation (SDE) models derived from continuous-time branching processes have been demonstrated to serve as valuable tools for the analysis of population dynamics (Allen 2015; Pardoux 2016). Specifically, in the context of epidemic dynamics, when the infectious population is of small magnitude relative to the overall population size, it becomes feasible to approximate the continuous-time dynamics using a nonlinear Stochastic Differential Equation (SDE) model, such as a stochastic logistic model that incorporates variations arising from birth and death processes (Allen 2015). In the realm of classical estimation methods, the challenges of devising procedures that strike a balance between computational efficiency and achieving optimal statistical performance have reached a well-established level of understanding.

Let us consider the parameter estimation problem of the following SDE driven by a standard strictly stable process $(Z_t)_{t \geq 0}$ defined in a given filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$:

$$\begin{cases} dX_t = f(X_t)dt + \rho g(X_{t-})dZ_t, & t \in [0, T], \\ X_0 = x_0, \end{cases} \quad (1)$$

where x_0 is a starting point, $T > 0$ stands for time horizon. Notably, in our investigation, we assume the knowledge of the function $g : \mathbb{R} \rightarrow \mathbb{R}$. The function $f : \mathbb{R} \rightarrow \mathbb{R}$, constant $\rho > 0$ and the stable process $(Z_t)_{t \in [0, T]}$ (which is considered as a nuisance parameter) are unknown. The random variable Z_1 which is a standard strictly stable distribution with parameters $\alpha \in (1, 2)$ (the index of stability) and $\beta \in [-1, 1]$ (the skewness parameter) will be defined in the sequel. The existence and uniqueness of solutions to the SDE (1) under Lipschitz conditions are standard results in stochastic calculus (Applebaum 2009). Nevertheless, we recall some findings. For $f = 0$ and g is Hölder continuous with exponent $\frac{1}{\alpha}$; if Z is a symmetric stable process then it is proved in Bass et al. (2004) that there exists a strong solution for $\alpha \in (1, 2)$. If f and g are continuous and have at most linear growth, then there is weak existence in L^p for any $p \in (0, \alpha)$, see Proposition 2 in Fournier (2013) for $\alpha \in (0, 2)$ and $\alpha \neq 1$. Moreover for $\alpha \in (1, 2)$ the pathwise uniqueness property holds whenever the function g is Hölder continuous with exponent lying in $[1 - \frac{1}{\alpha}, \frac{1}{\alpha}]$. When the drift term f is Hölder continuous of order lying in $(2 - \alpha, 1)$ and g is Lipschitz continuous and bounded, then existence and uniqueness of a strong solution is derived in Mikulevičius and Xu (2018) and Pamen and Taguchi (2017).

The case study of the parameter estimation theory for SDE by Brownian motions are well known in the literature. Nevertheless, one can mention some traditional

methods such as the maximum likelihood estimator (MLE), the least squares estimator (LSE) techniques (Lipcer et al. 2001; Dorogovcev 1976; Le Breton 2009; Craigmile et al. 2023); the consistency and the asymptotic distribution (Dorogovcev 1976; Le Breton 2009; Kasonga 1988; Rao 1983; Aït-Sahalia 2002). For further recent investigation, we refer to Kutoyants (2004) and the references therein. Significant advancements have been achieved in the realm of parameter estimation for Stochastic Differential Equations (SDEs) driven by Lévy processes with finite moments. The investigation conducted in Masuda (2005) centered on the consistency and asymptotic normality when the driving process manifests as a zero-mean adapted Lévy process with finite moments. The study of the asymptotic normality of the Least Squares Estimator (LSE) and Maximum Likelihood Estimator (MLE) for the pure jump case is extensively explored in Shimizu (2006); Shimizu and Yoshida (2006). It is noteworthy that research has been undertaken in instances where the driving process Z takes the form of a stable Lévy process, characterized by its distinctive infinite variance property. However, in this scenario, the MLE loses its validity as the explicit density function is not universally available, and the Girsanov measure transformation encounters challenges for α -stable processes with $\alpha \in (1, 2)$. Key contributions in this domain are highlighted in the following well-established results. In Hu and Long (2007), Hu and Long (2009), authors use the trajectory fitting method in conjunction with the weighted least squares technique for the drift coefficient of an α -stable driven Ornstein-Uhlenbeck (OU) process, where $\alpha \in (1, 2)$. This approach is applicable when the process is observed at discrete time instants, encompassing both ergodic and non-ergodic cases. The work also delves into the consistency and asymptotic distribution of the estimator, showcasing a higher order of convergence compared to the classical Gaussian case. The investigation conducted in Li and Ma (2015), Wei (2020) revolves around the drift parameter estimation of a stable driven Cox-Ingersoll-Ross (CIR) model, a special subcritical continuous-state branching process with immigration. The authors derive the consistency and central limit theorems of the conditional least squares estimators and the weighted conditional least squares estimators of the drift parameters based on low-frequency observations. Results in Dexheimer and Strauch (2022) explores Lasso and Slope drift estimators for Lévy-driven Ornstein-Uhlenbeck processes. Furthermore, Yang (2017) focuses on a maximum likelihood-type estimation of the drift and volatility constant coefficient parameters in a stable-driven CIR model. In a recent contribution (Bayraktar and Clément 2023), the author addresses the joint parameter estimation of the drift parameters, scaling parameter for the diffusion coefficient, and the stability index α for the jump activity parameter. This estimation is conducted from high-frequency observations of the stable CIR process over a fixed time period. The methodology employed is grounded in the approximation of the conditional distribution. Now, we shift our focus to non-parametric estimation, a pivotal domain in statistics that involves the estimation of an unknown function from a data sample without predefining a formula. Numerous authors have delved into the non-parametric estimation of the drift function, denoted as f , within the framework of diffusions driven by Brownian motions, considering various conditions. The fundamental statistical properties, including the consistency and rate of convergence, have been extensively investigated for non-parametric methods, particularly the

Nadaraya-Watson (N-W) estimators (Nadaraya 1964; Watson 1964). These investigations span scenarios under both independence with identically distributed data and weak dependence conditions, such as mixing conditions. A comprehensive overview of non-parametric methods for diffusion processes can be found in the survey paper authored by Iacus et al. (2008). In the stable and non-parametric context, additional insights are explored in Wu (2003), Long and Qian (2013), as well as Rao (2021), Zhang et al. (2019), Lin et al. (2014). However, tackling the estimation of the diffusion part (ρ , g and Z parameters) proves to be considerably more challenging (Long and Qian 2013).

In this paper, we operate under the assumption that the function g is known and provided. We consider real stochastic processes X observed at some discrete times, that is a solution of the SDE (1) driven by α -stable processes, $\alpha \in (1, 2)$. We assume that, the process X is observed at discrete time points $\{t_i = i\Delta_n, i = 0, 1, 2, \dots, n - 1\}$ with Δ_n a time frequency of the observation and n is the sample size. If the solution of the Stochastic Differential Equation (SDE) represented by (1) is stationary, and the stationary distribution possesses a continuous density and is strongly mixing, the authors in Long and Qian (2013) establish a non-parametric estimation of the drift function utilizing the Nadaraya-Watson (N-W) estimator. The statistical properties, such as consistency and the rate of convergence of the N-W estimators, under dependence conditions like mixing, are systematically developed in Long and Qian (2013). In instances where the drift function is linear, the estimation for both discrete and continuous observations has been extensively investigated in previous works, including (Hu and Long 2009, 2007; Li and Mytnik 2011; Bayraktar and Clément 2023; Yang 2017), encompassing stable-driven Cox-Ingersoll-Ross (CIR) and Ornstein-Uhlenbeck (OU) model types. All methods developed and applied within the framework of stable-driven CIR and OU models are conditionally dependent on the observation of the noise $Z = (Z_t)_{t \in [0, T]}$ parameter. Their performance is strongly influenced by the time frequency Δ_n and the time horizon T . While parameter estimation for stable-driven Stochastic Differential Equations (SDE's) has seen significant development in recent years, few works, such as Bayraktar and Clément (2023) in the CIR case, delve into joint parameter estimation for the drift coefficient, scaling parameter for the diffusion coefficient and, the jump activity diffusion parameters. These parameters include the stability index α and the scaling diffusion parameter ρ .

In this article, our primary objective is to address the estimation of diffusion parameters for discretely observed stable driven stochastic differential equations (1), assuming that only the function g is known. Our framework presupposes that the solution X is stationary and strongly mixing, with a connection emphasized with ergodic conditions. The proposed estimation procedure introduces a novel methodology that utilizes the N-W estimator for the drift function. Specifically, using the Euler-Maruyama scheme, we derive estimators for the parameters associated with the diffusion part (scaling parameter ρ and the stable noise parameters α and β) from a given sample characteristic function and employing linear (or weighted) regression techniques. Secondly, in cases where the drift function is linear, such as in stable-driven Cox-Ingersoll-Ross (CIR) and Ornstein-Uhlenbeck (OU) processes, we establish estimators for the drift coefficients using Itô formula and employing again

linear regression techniques, we derived unbiased and consistent estimators of their drift coefficients.

The organizational structure of this paper is delineated as follows. In Sect. 2, we revisit fundamental facts about stable processes and relevant assumptions for the classical N-W estimation framework for the drift function of the SDE (1). After that, we present our main theorems which, articulate explicit formulas for the estimators of the diffusion parameters (scaling parameter ρ and the stable noise parameters α and β). Additionally, we provide estimators for the drift coefficients in cases where the drift is linear, specifically for stable-driven CIR and OU processes. Section 3 is dedicated to presenting the proofs of our main results and Sect. 4 delves into the numerical performance evaluation of the estimators using synthetic data generated from mixing processes as well as the connection between ergodicity and mixing conditions for Markov processes that are solutions of SDE (1). We apply the best performance results to real financial data, such as the exchange rate of the Canadian dollar against the US dollar over a fixed period of time. Finally, in the Appendix section, we outline all material such as tables and figures.

2 Methodology and main results

In this section, our primary objectives are twofold. Firstly, we aim to introduce preliminaries tools in order to provide the reader with a comprehensive reading of the main results. We give a brief description of stable laws and processes, emphasizing key assumptions concerning a non-parametric estimation method of the drift function of the SDE (1) through of the the N-W estimator. Secondly, we provide the statements for our main results that involves the simultaneous estimation of the scaling parameter ρ and the parameters α and β characterizing the standard strictly stable process in the diffusion component of the SDE (1). We also focus on the estimation of common drift coefficients for some well-known stochastic models.

2.1 Preliminaries tools

Throughout this paper we assume that there is a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ satisfying the usual conditions, i.e.,

- (1) $(\Omega, \mathcal{F}_t, \mathbb{P})$ is complete for all $t \in \mathbb{R}_+$, \mathcal{F}_0 contains all the P-null sets in \mathcal{F} for all $t \in \mathbb{R}_+$.
- (2) $\mathcal{F}_t = \mathcal{F}_{t+}$ where $\mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s$, for all $t \geq 0$, i.e. the filtration is right-continuous.

Stable laws, introduced by Paul Lévy in 1925, emerge as the limit of normalized sums. The definitions and properties provided below are drawn from sources such as Samorodnitsky et al. (1996) and Nolan (2020). It's important to note that there exist multiple characterizations of stable distributions, each with distinct advantages depending on the intended applications. The following parameterization, denoted as parameterization 0, proves to be particularly useful for computer processing.

Definition 2.1 A random variable $X \sim S_\alpha(\gamma, \beta, \zeta)$ if,

$$\Psi(t) = \begin{cases} \exp\left(-\gamma^\alpha |t|^\alpha \left[1 + i\beta \left(\tan\left(\frac{\pi\alpha}{2}\right)\right) \text{sign}(t) (|\gamma t|^{1-\alpha} - 1)\right] + i\zeta t\right) & \text{if } \alpha \neq 1, \\ \exp(-\gamma |t| [1 + i\beta \frac{2}{\pi} \text{sign}(t) \log(\gamma |t|)] + i\zeta t) & \text{if } \alpha = 1, \end{cases}$$

where $\alpha \in (0, 2]$ is the index of stability, $\beta \in [-1, 1]$ the skewness parameter; $\gamma > 0$ the scale parameter and $\zeta \in \mathbb{R}$ the location or shift parameter and the sign function defined by:

$$\text{sign}(t) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

For a more in-depth exploration of stable distributions, please refer to Nolan (2020). When $\gamma = 1$, we designate X as a standard α -stable random variable. For $\gamma = 1$ and $\zeta = 0$, X is called a strictly α stable random variable, and in this case it is known that $\mathbb{E}(X) = 0$ whenever $\alpha \in (1, 2)$. In addition, if $\beta = 0$, it is characterised as symmetric. In general $\mathbb{E}(|X|^p) < \infty$ for any $p < \alpha$ and $\mathbb{E}(|X|^p) = +\infty$ for any $p \geq \alpha$, so the variance is infinite and in the case $\alpha \in (0, 1)$ the mean is infinite. Let's now introduce α -stable processes, which can be regarded as a special class of Lévy processes. For further insight, one can consult established literature such as Ken-Iti (1999).

Definition 2.2 An \mathcal{F}_t -adapted stochastic process $Z = \{Z_t\}_{t \geq 0}$ is called an α -stable process if

- (1) $Z_0 = 0$, a.s.;
- (2) Z has α -stable stationary increments distributions $Z_t - Z_s \sim Z_{t-s} \sim S_\alpha(\gamma(t-s)^{1/\alpha}, \beta, \zeta)$, $t > s \geq 0$.
- (3) For any time points $0 \leq s_0 < \dots < s_m < \infty$, the random variables $Z_{s_0}, Z_{s_1} - Z_{s_0}, \dots, Z_{s_m} - Z_{s_{m-1}}$ are independent.

For insights into the case of standard α -stable processes, Long and Qian (2013) is a recommended reference. Set $\Delta Z_t = Z_t - Z_{t-}$. The jumps part of the stochastic process Z can be described by its Poisson random measure (jump measure of Z on interval $[0, t]$) defined as $N(t, A) = \sum_{0 \leq s \leq t} \mathbb{1}_A(\Delta Z_s)$, $A \in \mathcal{B}(\mathbb{R}^*)$, the number of jumps of Z on the interval $[0, t]$ whose size lies in the set A bounded below, that is, A admits a lower bound (in \mathbb{R}). For such A , the process $N(\cdot, A)$ is a Poisson process and the Lévy measure $\nu(A) := \mathbb{E}(N(1, A))$ defined on \mathbb{R}^* has the following explicit form: $\nu(dx) := \frac{dx}{|x|^{\alpha+1}} (c_+ 1_{\{x>0\}} + c_- 1_{\{x<0\}})$. In the sequel, we'll denote by $\tilde{N}(t, A)$ the compensated Poisson measure. Note that in the case $\alpha \in (1, 2)$, the characteristic function of a strictly α -stable process is reduced into the form:

$$\Psi_{Z_t}(u) = \exp t \left(\int_{-\infty}^{+\infty} (e^{iuy} - 1 - iuy) \nu(dy) \right), \quad t \geq 0.$$

The parameters c_+ , c_- mentioned above are non-negative, with the additional condition that $c_+ + c_- > 0$. There exists a connection between jump measurement coefficients c_+ , c_- and the skewness parameter, for instance, $\beta = \frac{c_+ - c_-}{c_+ + c_-}$. If $\beta > 0$, we will say that Z admits a positive jump activity. For numerical application, we use a random walk approximation method based on the work of Chambers et al. (1976), Janicki et al. (1997), proves useful to simulate such a stable process. Naturally, other methods are also available, such as the series approximation of Lévy processes, see Janicki et al. (1997) for more details.

In this paper, we first consider the non-parametric estimation framework for the drift function of the above class of stable driven SDE equations:

$$X_t = x_0 + \int_0^t f(X_s) ds + \rho \int_0^t g(X_{s-}) dZ_s, \quad t \in [0, T],$$

where Z is a α -stable process with $\alpha \in (1, 2)$, not necessarily symmetric. We consider several standard sufficient conditions on f and g that implies the existence of a unique strong solution. As far as applications are concerned, we will look at examples for which strong existence is well known. Recall that, under linear growth assumptions, there is a weak solution, see Fournier (2013), Proposition 2. In particular, weak solution conditions covers the case where f and g are continuous Lipschitz or Hölder functions. The author establishes weak existence through the Picard approximation. In some cases the weak solution is adapted and unique, see for example Priola (2012) and Mikulevičius and Xu (2018) as already mentioned. Suppose that g is non-decreasing and Hölder continuous with index $1 - \frac{1}{\alpha}$ and f is the sum of a Lipschitz continuous function and a non-increasing function. It is shown in Li and Mytnik (2011), Fu and Li (2010) that both the strong existence and path uniqueness hold if $\alpha \in (1, 2)$ and Z has only positive jumps. The Nadaraya-Watson (N-W) estimator is a classical method for estimating the drift function f . The mixing property of the solution is thus essential in the framework of the N-W estimation. Let us recall some conventional mixing coefficients.

Definition 2.3 (Mixing conditions, see Manou-Abi 2015) Let \mathcal{F}_s (resp. \mathcal{G}_s) be respectively the backward (or the past) and the forward (or the future) σ -fields generated by X_u for $0 \leq u \leq s$ (resp. $u \geq s$).

1. The strong mixing coefficient $\alpha_{mix}(t)$ is defined as:

$$\begin{aligned} \alpha_{mix}(t) &= \sup_s \sup_{A \in \mathcal{F}_s, B \in \mathcal{G}_{s+t}} |\mathbb{P}(A \cup B) - \mathbb{P}(A)\mathbb{P}(B)| \\ &= \frac{1}{4} \sup_s \left\{ \sup_{F, G} \text{Cov}(F, G), F \mathcal{F}_s(\text{resp. } G \mathcal{G}_{s+t}) \text{ measurable and bounded by 1.} \right\} \end{aligned}$$

If $\lim_{t \rightarrow \infty} \alpha_{mix}(t) = 0$, the process is strongly mixing.

2. The β -mixing coefficient $\varphi(t)$ is defined as:

$$\beta_{mix}(t) = \sup_s \left\{ \sup_{A,B} (\mathbb{P}(B|A) - \mathbb{P}(B)), A \in \mathcal{F}_s, B \in \mathcal{G}_{s+t} \right\}.$$

If $\lim_{t \rightarrow \infty} \beta_{mix}(t) = 0$, the process is β -mixing or uniformly mixing.

3. The ρ -mixing coefficient $\rho_{mix}(t)$ is defined as the maximal correlation coefficient, i.e.

$$\rho_{mix}(t) = \sup_s \left\{ \sup_{F,G} \text{Corr}(F, G), F \in L^2(\mathcal{F}_s), G \in L^2(\mathcal{G}_{t+s}) \right\}.$$

If $\lim_{t \rightarrow \infty} \rho_{mix}(t) = 0$ the process is ρ -mixing.

Please, note that if X is a stationary process (i.e. the law of X_{t+s} is the same as the one of X_s), the supremum on s is irrelevant. The process X is geometrically (or exponentially) strong mixing if $\alpha_{mix}(t) \leq c_0 a^t$ with $a \in (0, 1)$ and $t \geq 0$. In a non-parametric estimation, we are concerned with a kernel function $K(\cdot)$, which is a symmetric and non-negative probability density function satisfying $\sup(1 \vee |u|)K(u) < M_0 < +\infty$ and

$$\int_{-\infty}^{+\infty} u^2 K(u) du < +\infty, \quad \int_{-\infty}^{+\infty} K^2(u) du < +\infty.$$

In the sequel, we make the following assumptions following (Long and Qian 2013). Similar assumptions are also presented in Lin et al. (2014) where authors study the local polynomial estimation under regular conditions.

(A₁) Suppose that the unique strong solution X_t is stationary and admits a unique invariant distribution π and is geometrically strongly mixing.

(A₂) Suppose that the density function $p(x)$ of the stationary distribution π is continuous.

(A₃). As $n \rightarrow \infty$: $h \rightarrow 0$, $\Delta_n \rightarrow 0$,

$$n\Delta_n \rightarrow \infty, \text{ and } \frac{n\Delta_n h}{(\log(n\Delta_n))^2} \rightarrow +\infty.$$

As already mentioned, the process X is observed at some discrete times $\{t_k = k\Delta_n, k = 0, 1, 2, \dots, n-1\}$ with Δ_n a time frequency of the observation and n is the sample size. The Euler-Maruyuma scheme of the above SDE is written as follows.

$$X_{t_{k+1}}^n = X_{t_k}^n + f(X_{t_k})\Delta_n + \rho g(X_{t_k})\Delta Z_k, \quad X_0^n = x \in \mathbb{R},$$

where $\Delta Z_k = Z_{t_{k+1}}^n - Z_{t_k}^n$. Let us first state some well known results about the consistency (efficacy) of the Euler-Maruyuma scheme which is a well-known method for

approximating any solution as well as Picard iteration. Let’s recall some findings results on the existence of a strong approximate solution. In Pamen and Taguchi (2017), when $g = 1$ and the drift function f is a bounded β -Hölder continuous in space with $\beta \in (0, 1)$, a strong approximate solution is proved and the rate of convergence for the Euler-Maruyama approximation is given. Similarly, the authors in Mikulevičius and Xu (2018) find a strong approximate solution and a rate of convergence of order $n^{-p\beta/\alpha}$ when $p \in (0, \alpha/\beta)$ in the case where Z is a symmetric stable process of index $\alpha \in [1, 2]$, f is a β -Hölder continuous function with $\beta > 1 - \frac{\alpha}{2}$ and g is a bounded Lipschitz continuous function. Note that, under the above linear growth conditions (C1) and (C2), a recent result on approximate solutions of the above SDE has been established in Manou-Abi (2023) for the above SDE. Set $Y_k = X_{t_{k+1}}^n - X_{t_k}^n$. The main idea of N-W estimator is to minimize the following object function: $\sum_{k=0}^{n-1} W_{n,k}(x)(Y_k - b\Delta_n)^2$, over the parameter space of b with certain weights functions $W_{n,k}(x)$ given by:

$$W_{n,k}(x) = \frac{K_h\left(\frac{X_{t_k}^n - x}{h}\right)}{\sum_{k=0}^{n-1} K_h\left(\frac{X_{t_k}^n - x}{h}\right)},$$

for all $x \in \mathbb{R}$, where $K_h(x) = (K(x/h))/h$ with K the Kernel and h the bandwidth parameter. Thus the N-W estimator of the drift function f is given by the following expression:

$$\hat{f}_n(x) = \sum_{k=0}^{n-1} W_{n,k}(x)Y_k, \quad \forall x \in \mathbb{R},$$

so that $\hat{f}_n(x) = \sum_{i=1}^n \frac{K_h\left(\frac{X_{t_i} - x}{h}\right)(X_{t_i} - X_{t_{i-1}})}{n^{-1} \sum_{i=1}^n K_h\left(\frac{X_{t_i} - x}{h}\right)}$ for all $x \in \mathbb{R}$. In this study, we inves-

tigate several kernel functions; however, we specifically adopt the Gaussian kernel function for simulation purposes: $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$.

2.2 Main results

Using the N-W estimator as a foundation, and the Euler-Maruyama scheme, the primary goal in this section is twofold. Firstly, it involves the simultaneous estimation of the scaling parameter ρ and the parameters α and β characterizing the standard strictly stable process in the diffusion component of the SDE (1). Secondly, we focus on the estimation of common drift coefficients θ and μ for the Stable driven CIR and OU processes in Eqs. (4) or (6). In the following, by a solution we mean a unique, strong solution.

Theorem 2.1 Let $X = (X_t)_{t \in [0,1]}$ be a real valued stationary and strongly mixing process with unique invariant distribution which is solution of the SDE (1). We assume that X is observed on a sample of discrete time points t_i with step Δ_n of size n such $g(X_{t_i}) \neq 0$. Assume that assumptions (A_2) and (A_3) holds. Let $(\hat{f}_n(X_{t_i}))_{i=1}^n$ be the sequence of the N-W estimate of the drift coefficient of the SDE (1). We define the following sample characteristic function

$$\hat{\varphi}_n(z_k) = \frac{1}{n} \sum_{i=0}^{n-1} \exp \left(jz_k \left(\frac{\Delta X_{t_i} - \hat{f}_n(X_{t_i}) \Delta_n}{g(X_{t_i})} \right) \right), \tag{2}$$

for all $k \in [[1, m]]$, $m > 0$, and $j^2 = -1$. Set

$$\left\{ \begin{array}{l} \hat{\alpha}_m = \frac{\sum_{k=1}^m W_k V_k - \frac{1}{m} \sum_{k=1}^m V_k \sum_{k=1}^m W_k}{\left(\sum_{k=1}^m W_k^2 \right)^{-\frac{1}{m}} \left(\sum_{k=1}^m W_k \right)^2}, \\ \hat{\lambda}_m = \frac{1}{m} \sum_{k=1}^m V_k - \frac{\hat{\alpha}_m}{m} \sum_{k=1}^m W_k, \\ \hat{\rho}_m = \Delta_n^{-\frac{1}{\hat{\alpha}_m}} \exp \frac{\hat{\lambda}_m}{\hat{\alpha}_m}, \\ \hat{\beta}_m = \frac{\sum_{k=1}^m z_k S_k}{\sum_{k=1}^m z_k B_k}, \end{array} \right. \tag{3}$$

where for all n large enough

$$S_k \sim \arg(\hat{\varphi}_n(z_k)), \quad B_k \sim \tan\left(\frac{\pi \hat{\alpha}}{2}\right) \text{sign}(z_k) (z_k - z_k^{\hat{\alpha}_m}) \Delta_n^{1/\hat{\alpha}_m},$$

and

$$V_k \sim \log(-\log |\hat{\varphi}_n(z_k)|) \quad \text{and} \quad W_k = \log(|z_k|).$$

For n large enough and good choice of m values of z ; $\hat{\alpha}_m$, $\hat{\rho}_m$ and $\hat{\beta}_m$ are least squares estimators of α , ρ and β .

If we consider some weights p_k , for instance $p_k = \frac{1}{\delta_k^2}$ the inverse of the variance of the k -th observation, a weighted least squares method lead to the following estimators:

$$\left\{ \begin{array}{l} \hat{\alpha}_m = \frac{\sum_{k=1}^m p_k W_k V_k - \frac{1}{\sum_{k=1}^m p_k} \sum_{k=1}^m V_k \sum_{k=1}^m W_k}{\left(\sum_{k=1}^m p_k W_k^2 \right)^{-\frac{1}{m}} \left(\sum_{k=1}^m p_k \right) \left(\sum_{k=1}^m W_k \right)^2}, \\ \text{Set } \hat{\lambda}_m = \frac{1}{\sum_{k=1}^m p_k} \sum_{k=1}^m p_k V_k - \frac{\hat{\alpha}_m}{\sum_{k=1}^m p_k} \sum_{k=1}^m p_k W_k, \\ \hat{\beta}_m = \frac{\sum_{k=1}^m p_k z_k S_k}{\sum_{k=1}^m p_k z_k B_k}. \end{array} \right.$$

Let's discuss about some formulations in order to check some consistency properties of these linear least squares estimators through a linear statistical regression model. Since (V_k) is a random sequence and $|\hat{\phi}_n(z_k)| \in [0, 1]$ one may say that $V = (V_k)_{k=1}^m$ for m well-chosen values of z_k can be obtained from a truncated uniform random variable U so that $-\log(U)$ follows a truncated exponential random variable T where occurrences is limited to finite positive values $[0, A]$ with $A > 0$. Hence V is a truncated distribution of $\log(T)$ such that $\mathbb{E}(\log(T)|_{[0,A]})$ and $\text{Var}(\log(T)|_{[0,A]})$ are finite. When the number of observations m is sufficiently large, a Gaussian distribution can be assumed for the error $\epsilon_k = V_k - \alpha \log(|z_k|) - \log(\rho^\alpha \Delta_n)$ that can be considered to be uncorrelated with the (deterministic) regressors $\log(|z_k|)$. Hence the above estimators become Maximum Likelihood Estimators. When the normality assumption is no longer valid, we can try to reduce the data to Gaussian distributions by means of data transformations, the symmetrization and standardization (since $\mathbb{E}(\log(T)|_{[0,A]})$ and $\text{Var}(\log(T)|_{[0,A]})$ are finite) offers an advantage. The Min-Max Scaling can be applied since the data may vary in different scales, reducing the effect of outliers.

The optimal value of m is suggested in Kogon and Williams (1998), advocating the selection of points z_k within the interval $[0.1, 1]$.

Theorem 2.1 applies particularly to the following α -stable-driven Ornstein Uhlenbeck (OU), Cox-Ingersoll-Ross (CIR) and Lotka-Volterra processes. Such models are popular in stochastic modelling for description of interest rates in finance and population dynamics. Assume that Z is a standard strictly stable process with parameters $\alpha \in (1, 2)$ and $\beta \in [-1, 1]$. A generalized Ornstein-Uhlenbeck (OU) process driven by a strictly standard α -stable process $(Z_t)_{t \geq 0}$ is defined to be the solution to the following linear stochastic differential equation

Definition 2.4 (Stable OU process)

$$dX_t = \theta(\mu - X_t)dt + \rho dZ_t, \quad X_0 = x \in \mathbb{R}, \quad (4)$$

where θ and μ are constants.

We can apply Theorem 2.1 to this model to estimate parameters ρ , α and β .

Definition 2.5 (Stable CIR process) A stable driven Cox-Ingersoll-Ross (Stable CIR Yang 2017) is defined by:

$$dX_t = \theta(\mu - X_t)dt + \rho|X_t|^{1/q}dZ_t, \quad X_0 \geq 0, \quad (5)$$

where $q > 0$, $\theta > 0$, $\rho > 0$, and $\mu \in \mathbb{R}$ are constants and Z is for example a strictly standard α -stable process with positive jump activity.

Note that if Z is symmetric with $\alpha \in (1, 2)$ and since $x \rightarrow |x|^{1/q}$ is Hölder continuous for $q \geq 1$, there exists a solution according to Bass et al. (2004). For $q = 2$ the Stable CIR is studied in Wei (2020) for symmetric stable process. According to Li and Mytnik (2011), there is a pathwise unique positive strong solution for

the above Stable CIR as $q^{-1} + \alpha^{-1} \geq 1$ when ρ is small whenever the process Z have only positive jumps. We can apply Theorem 2.1 to this model to estimate parameters ρ , α and β when the parameter q is known.

For $q = \alpha \in (1, 2)$ and in the case where Z is a pure-jump α -stable Lévy process Z with positive jumps the following stable driven CIR is introduced in the literature (see Bayraktar and Clément 2023 or Li and Ma 2015) as a particular form of the continuous-state branching processes with immigration, which emerge as scaling limits of Galton-Watson branching processes with immigration (CBI-processes, Li et al. 2022; Li and Ma 2015 and Pardoux 2016).

Definition 2.6 (SCIR process) The SCIR process is defined by:

$$dX_t = \theta(\mu - X_t)dt + \rho|X_t|^{1/\alpha}dZ_t, \quad X_0 \geq 0, \quad (6)$$

where $\theta > 0$, $\rho > 0$, and $\mu \in \mathbb{R}$ are constants and Z is a standard and positive strictly α -stable process.

It is shown whenever that the solution is positive that is $\mathbb{P}(X_t > 0) = 1$, whenever $\mu > 0$, $\theta > 0$, $\rho > 0$ and $x_0 > 0$. Unfortunately, we cannot estimate α with this model, as q is a function of α . Only the parameters ρ and β must be unknown to apply Theorem 2.1.

The following stable driven Lotka and Volterra extension was studied in Zhang et al. (2017) for $q = 1$:

Definition 2.7 (A Stable driven Lotka-Volterra process) A Stable driven Lotka-Volterra process can be defined as follows

$$dX_t = X_t(\lambda - \theta X_t)dt + \rho|X_t|^{1/q}dZ_t, \quad X_0 \geq 0, \quad (7)$$

where λ , ρ , θ are real constants.

The author proved in Zhang et al. (2017) for $q = 1$, the existence of a unique global positive solution of the above stable driven Lotka-Volterra process, if Z is a spectrally positive strictly α -stable process for any $\alpha \in (0, 2)$. Here again, we can apply Theorem 2.1 to this model to estimate parameters ρ , α and β when the parameter q is known.

In what follows, we discuss the parameter estimation of the drift coefficients θ and μ of the above stable driven OU and the SCIR processes.

Theorem 2.2 Let $X = (X_t)_{t \in [0, T]}$ be a real valued stationary and strongly mixing (ergodic) process satisfying (4) or (6). We assume that X is observed on an n -sample of discrete time points t_i with fixed time frequency Δ . Assume that assumptions (A_2) and (A_3) holds. Let $(\hat{f}_n(X_{t_i}))_{i=1}^n$ be the sequence of the N-W estimate of the linear drift function in (4) or (6). Set

$$\begin{cases} \hat{a}_n = \frac{\sum_{i=0}^{n-1} X_{t_i} \hat{f}_n(X_{t_{i+1}}) - \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i} \sum_{i=1}^n \hat{f}_n(X_{t_{i+1}})}{\frac{1}{n} \left(\sum_{i=0}^{n-1} X_{t_i} \right)^2 - \sum_{i=0}^{n-1} X_{t_i}^2}, \\ \hat{\mu}_n = \frac{1}{n} \sum_{i=0}^{n-1} \hat{f}_n(X_{t_{i+1}}) - \hat{a}_n \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i}, \\ \hat{\theta}_n = -\frac{1}{\Delta} W(-\hat{a}_n \Delta), \end{cases} \tag{8}$$

where W is the Lambert function, see (Corless et al. 1996).

- We assume moreover that $\alpha \in (\sqrt{2}, 2)$ for the driving process Z in the SCIR process (6).

For n large enough, $\hat{\theta}_n$ and $\hat{\mu}_n$ are unbiased and consistent estimators of θ and μ .

3 Proofs of the main theorems

In this section, we provide the proofs for our main results.

3.1 Proof of theorem 2.1

Proof We begin by revisiting the Euler approximation of the SDE (1):

$$X_{t_{i+1}}^n = X_{t_i}^n + f(X_{t_i}^n) \Delta_n + g(X_{t_i}^n) Y_i^n,$$

where $Y_i^n = \rho \Delta Z_i^n = \rho(Z_{i+1}^n - Z_i^n)$, $i = 0, \dots, n - 1$, are identically sequence distributed as $Z_{\Delta_n} \sim S_\alpha(\Delta_n^{1/\alpha} \rho, \beta, 0)$. We have

$$Y_i^n = \frac{\Delta X_{t_i} - \hat{f}_n(X_{t_i}) \Delta_n}{g(X_{t_i})}, \quad \forall i = 0, \dots, n - 1.$$

We can define a sample characteristic function as in (2):

$$\hat{\varphi}_n(u) = \frac{1}{n} \sum_{i=0}^{n-1} \exp\left\{ju \left(\frac{\Delta X_{t_i} - \hat{f}_n(X_{t_i}) \Delta_n}{g(X_{t_i})} \right)\right\}, \quad \forall u \in \mathbb{R},$$

which is an asymptotic approximation of the characteristic function $\varphi_{Z_{\Delta_n}}(u)$ of Z_{Δ_n} for n large enough Now, using Definition 2.1 or 2.2 and, utilizing the following formula:

$$\log(-\log |\varphi_{Z_{\Delta_n}}(u)|) = \log(\rho^\alpha \Delta_n) + \alpha \log(|u|),$$

we can establish an ordinary least squares regression method to estimate α and ρ . To achieve this, it is desired to find the vector parameter (α, λ) such that the underlined linear function fits best the given data in the least squares sense, that is, the following sum of squares is minimized:

$$G(\lambda, \alpha) = \sum_{k=1}^m (V_k - \lambda - \alpha W_k)^2,$$

$$(\hat{\lambda}, \hat{\alpha}) = \arg \min_{(\lambda, \alpha)} G(\lambda, \alpha),$$

where

$$V_k \sim \log(-\log |\hat{\varphi}_n(z_k)|), \quad W_k = \log(|z_k|), \quad \lambda \sim \log(\rho^\alpha \Delta_n).$$

The minimum value of $G(\lambda, \alpha)$ occurs when the gradient is zero. Since the model contains two parameters, there are two gradient equations:

$$\begin{cases} \frac{\partial G(\lambda, \alpha)}{\partial \alpha} = -2 \sum_{k=1}^m W_k (V_k - \lambda - \alpha W_k), \\ \frac{\partial G(\lambda, \alpha)}{\partial \lambda} = -2 \sum_{k=1}^m (V_k - \lambda - \alpha W_k), \end{cases}$$

and these gradient equations have a closed solution given by:

$$\begin{cases} \frac{\partial G(\lambda, \alpha)}{\partial \alpha} = 0 \\ \frac{\partial G(\lambda, \alpha)}{\partial \lambda} = 0 \end{cases} \implies \begin{cases} \hat{\alpha}_m = \frac{\sum_{k=1}^m W_k V_k - \frac{1}{m} \sum_{k=1}^m V_k \sum_{k=1}^m W_k}{\sum_{k=1}^m W_k^2 - \frac{1}{m} \left(\sum_{k=1}^m W_k \right)^2}, \\ \hat{\lambda}_m = \frac{1}{m} \sum_{k=1}^m V_k - \frac{\hat{\alpha}_m}{m} \sum_{k=1}^m W_k. \end{cases}$$

To estimate β , we use Definition 2.2 and the following formula:

$$S_k = \beta B_k + \zeta u_k \quad \text{here we have, } \zeta = 0,$$

where

$$S_k \sim \arg(\hat{\varphi}_n(u_k)), \quad B_k \sim \tan\left(\frac{\pi \hat{\alpha}}{2}\right) \text{sign}(u_i) (u_k - u_k^{\hat{\alpha}}) \Delta_n^{1/\hat{\alpha}}.$$

Note that we set $\zeta = 0$ since we consider in this work strictly α -stable process Z . We employ once again the aforementioned least squares method to minimize the following sum of squares:

$$T(\beta, \zeta) = \sum_{k=1}^m (S_k - \beta B_k - \zeta u_k)^2.$$

From the gradient equations:

$$\begin{cases} \frac{\partial T(\beta, \zeta)}{\partial \beta} = -2 \sum_{k=1}^m B_k (S_k - \zeta u_k - \beta B_k), \\ \frac{\partial T(\beta, \zeta)}{\partial \zeta} = -2 \sum_{k=1}^m u_k (S_k - \zeta u_k - \beta B_k), \end{cases}$$

we have the following closed solution when $\zeta = 0$:

$$\hat{\beta}_m = \frac{\sum_{k=1}^m u_k S_k}{\sum_{k=1}^m u_k B_k}.$$

□

3.2 Proof of theorem 2.2

Proof We have

$$\hat{f}(X_{t_{i+1}}) = \theta(\mu - X_{t_{i+1}}).$$

If X satisfy (4) we have the following integral representation:

$$X_t = e^{-\theta t} X_0 + \theta \mu \int_0^t e^{-\theta(t-s)} ds + \rho \int_0^t e^{-\theta(t-s)} dZ_s, \quad t \geq 0.$$

so that for any $t \geq r \geq 0$:

$$X_t = e^{-\theta(t-r)} X_r + \theta \mu \int_r^t e^{-\theta(t-s)} ds + \rho \int_r^t e^{-\theta(t-s)} dZ_s, \quad t \geq 0.$$

We refer to Ken-Iti (1999), Chapter 3 for the case when $\mu = 0$ and to Hu and Long (2007) for the case $\mu \neq 0$. If X satisfy (6), note that for $\alpha \in (1, 2)$ it is well known that the positive stable process Z admits the following representation:

$$Z_t = \int_0^t \int_0^{+\infty} z \tilde{N}(ds, dz), \quad t \geq 0.$$

and applying Itô’s formula for Jump processes (Privault 2016), we have the following integral representation (see also Li and Ma 2015):

$$X_t = e^{-\theta(t-r)} X_r + \theta \mu \int_r^t e^{-\theta(t-s)} ds + \rho \int_r^t e^{-\theta(t-s)} |X_{s-}|^{1/\alpha} dZ_s, \quad t \geq 0.$$

We have

$$X_{t_i} = e^{-\theta(t_i-t_{i-1})} X_{t_{i-1}} + \theta \mu \int_{t_{i-1}}^{t_i} e^{-\theta(t_i-s)} ds + \epsilon_i,$$

where

$$\epsilon_i = \begin{cases} \rho \int_{t_{i-1}}^{t_i} e^{-\theta(t_i-s)} |X_{s-}|^{1/\alpha} dZ_s & \text{if } X \text{ is the stable CIR process in (6)} \\ \rho \int_{t_{i-1}}^{t_i} e^{-\theta(t_i-s)} dZ_s & \text{if } X \text{ is the stable OU process in (4).} \end{cases}$$

Note that if X is the stable OU process in (4), by basic properties of α -stable stochastic integrals for deterministic integrands (Taqqu 1994; Rosinski and Woyczynski 1986), the sequence (ϵ_i) is a α -stable random variable with distribution $S_\alpha(\rho\tau_n^{1/\alpha}, \beta, 0)$ with $\tau_n = \frac{1-e^{-\theta\Delta n}}{\theta\alpha}$ so that it is centered random sequence (also martingales difference). In the case where X is SCIR process in (6), according to the martingale property of a compensated Poisson stochastic integral it is a martingale differences if $\mathbb{E}|X_s|^{2/\alpha}$ is finite; which is the case whenever $\frac{2}{\alpha} < \alpha$ i.e. $\alpha \in (\sqrt{2}, 2)$. Therefore,

$$X_{t_i} = e^{-\theta\Delta}X_{t_{i-1}} + \mu(1 - e^{-\theta\Delta}) + \epsilon_i,$$

where (ϵ_i) is a centered random sequence (martingale differences). Finally,

$$\hat{f}(X_{t_{i+1}}) = \theta(\mu - X_{t_{i+1}}) = a\mu - aX_{t_i} + \eta_i, \quad a = \theta e^{-\theta\Delta}, \quad i \geq 0,$$

where $(\eta_i)_{i \geq 0}$ is again a centered random sequence (martingale differences). We can set up a Linear Least Square Method which consists in minimizing the following objective function:

$$\mathbb{G}(a, \mu) = \sum_{i=0}^{n-1} (\hat{f}(X_{t_{i+1}}) - a\mu + aX_{t_i})^2.$$

Using, the gradient equations:

$$\begin{cases} \frac{\partial \mathbb{G}(a, \mu)}{\partial a} = 2(-\mu + X_{t_i}) \sum_{i=0}^{n-1} (\hat{f}(X_{t_{i+1}}) - a\mu + aX_{t_i}), \\ \frac{\partial \mathbb{G}(a, \mu)}{\partial \mu} = -2a \sum_{i=0}^{n-1} (\hat{f}(X_{t_{i+1}}) - a\mu + aX_{t_i}), \end{cases}$$

it is straightforward to obtain the above already mentioned least square estimators:

$$\begin{cases} \hat{a}_n = \frac{\sum_{i=0}^{n-1} X_{t_i} \hat{f}_n(X_{t_{i+1}}) - \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i} \sum_{i=1}^n \hat{f}_n(X_{t_{i+1}})}{\frac{1}{n} (\sum_{i=0}^{n-1} X_{t_i})^2 - \sum_{i=0}^{n-1} X_{t_i}^2}, \\ \hat{\mu}_n = \hat{a}_n^{-1} \frac{1}{n} \sum_{i=0}^{n-1} \hat{f}_n(X_{t_{i+1}}) + \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i}, \\ \hat{\theta}_n = -\frac{1}{\Delta} W(-\hat{a}_n \Delta), \end{cases}$$

where W is the Lambert function. Now, let us prove the consistency of these estimators. Set $U_i = \hat{f}_n(X_{t_{i+1}})$ and note that since we assume that the process X is ergodic

and $\alpha \in (1, 2)$ then $\bar{X} = \frac{1}{n} \sum_{i=0}^{n-1} X_{t_i}$ converge to μ and $\bar{U} = \frac{1}{n} \sum_{i=0}^{n-1} U_i$ also converge. We have

$$\hat{a}_n = \frac{\sum_{i=1}^n X_{t_i} U_i - \frac{1}{n} \sum_{i=1}^n U_i \sum_{i=1}^n X_{t_i}}{\frac{1}{n} \sum_{i=1}^n X_{t_i} \sum_{i=1}^n X_{t_i} - \sum_{i=1}^n X_{t_i}^2} = \frac{\sum_{i=1}^n X_{t_i} (U_i - \bar{U})}{\sum_{i=1}^n X_{t_i} (\bar{X} - X_{t_i})},$$

where $\bar{U} = \frac{1}{n} \sum_{i=1}^n U_i$ and $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_{t_i}$,

so that

$$= \frac{\sum_{i=1}^n (X_{t_i} - \bar{X})(\bar{U} - U_i)}{\sum_{i=1}^n (X_{t_i} - \bar{X})^2} \text{ since } \sum_{i=1}^n \bar{X}(U_i - \bar{U}) = 0 \text{ and } \sum_{i=1}^n \bar{X}(X_{t_i} - \bar{X}) = 0.$$

From the relation

$$U_i - \bar{U} = a(\bar{X} - X_{t_i}) + \varepsilon_i,$$

where (ε_i) is again a centered random sequence, we have

$$\hat{a}_n = \frac{\sum_{i=1}^n (X_{t_i} - \bar{X})(a(X_{t_i} - \bar{X}) - \varepsilon_i)}{\sum_{i=1}^n (\bar{X} - X_{t_i})^2} = a + \frac{\sum_{i=1}^n (X_{t_i} - \bar{X})\varepsilon_i}{\sum_{i=1}^n (\bar{X} - X_{t_i})^2} \xrightarrow{n \rightarrow +\infty} 0,$$

since the process $X = \{X_t, t \geq 0\}$ is ergodic, stationary and have infinite square variation. We deduce the consistency of $\hat{\theta}_n$ and $\hat{\mu}_n$ by the use of continuous mapping theorem. □

4 Examples of mixing processes and numerical application

In this section, we present firstly, the link between ergodicity and mixing conditions for Markov processes that are solutions of SDE driven by Lévy stable processes. Secondly we discuss mixing conditions of the above mentioned stochastic models. We end by the numerical performance of our estimation procedure on synthetic data and an application to real data. Let $(X_t, t \geq 0)$ be an ergodic Markov process with unique invariant measure π , solution of a SDE driven by Lévy processes. Consider the Markov semigroup $(P_t)_{t \geq 0}$ associated to $(X_t, t \geq 0)$ and defined by $P_t \phi(x) = \mathbb{E}(\phi(X_t) / X_0 = x)$ for all ϕ in $L^p(\pi)$ or measurable and positive functions ϕ . Recall that P is a bounded operator in all $L^p(\pi)$, $p \geq 1$ with operator norm equal to 1 (i.e. a contraction). The adjoint operator P^* is defined by $\int \phi P_t^* f d\pi = \int f P_t \phi d\pi$. for functions ϕ and f that are square integrable with respect to π , this operator is once more a contraction. The subsequent definition introduces a method for regulating the ergodic decay to equilibrium.

Definition 4.1 (Ergodic rates of convergence Cattiaux and Manou-Abi 2014) For any $r \geq p \geq 1$ and $t \geq 0$ we define the following ergodic rates

$$\eta_{p,r}(t) = \sup_{\|\phi\|_{L^r(\pi)} \int \phi d\pi = 0} \|P_t \phi\|_{L^p(\pi)}.$$

The process X is said to be uniformly ergodic if $\lim_{t \rightarrow +\infty} \eta_{2,\infty}(t) = 0$.

The following definitions can be found in Zhang et al. (2017).

Definition 4.2 (Exponentially or strongly ergodic process) Assume that $X = (X_t, t \geq 0)$ is an ergodic Markov process with unique invariant measure π and $X_0 = x$.

1. The process X with is called exponentially ergodic if there exist a constant $k > 0$ and a positive measurable function $c(x)$ such that for all $t > 0$ we have

$$\|P_t(x, \cdot) - \pi\|_{var} \leq c(x)e^{-kt}.$$

2. The process X with is called strongly ergodic if there exist two constants $k, C > 0$ such that for all $t > 0$ we have

$$\|P_t(x, \cdot) - \pi\|_{var} \leq Ce^{-kt},$$

where $\|\cdot\|_{var}$ denotes the total variation norm on the space of signed probability measures defined by

$$\begin{aligned} \|P_t(x, \cdot) - \pi\|_{var} &= \sup_{A \in \mathcal{F}} |P_t(x, A) - \mu(A)| \\ &= \sup_{\|\phi\|_{\infty} \leq 1 \text{ and } \text{Law}(Y) = \pi} |\mathbb{E}\phi(X_t) - \mathbb{E}\phi(Y)|. \end{aligned}$$

From this definition, one can state that the process X is said to be strongly or exponentially ergodic if $\lim_{t \rightarrow +\infty} \eta_{1,\infty}(t) = 0$, where $\eta_{1,\infty}$ is termed the strong or exponential ergodic decay rate. The following lemma, as contained in Cattiaux and Manou-Abi (2014), enables a connection between ergodicity and mixing conditions.

Lemma 4.1 For all $t \geq 0$, we have

- (1) $\eta_{\infty,2}^2(t) \vee (\eta^*)_{\infty,2}^2(t) \leq \alpha_{mix}(t) \leq \eta_{\infty,2}([t/2]) \eta_{\infty,2}^*(t/2)$.
- (2) Either $\eta_{2,2}(t) = 1$ for all t or $\eta_{2,2}(t) \leq c e^{-\lambda t}$ for some $\lambda > 0$. In the second case $\eta_{2,2}^2(t) = (\eta^*)_{2,2}^2(t) \leq \rho_{mix}(t) \leq c \eta_{2,2}(t)$.
- (3) $\beta_{mix}(t) \leq \eta_{1,\infty}^2(t/2)$.

From the above Lemma one easily derive the following result.

Theorem 4.1 Assume that $X = (X_t, t \geq 0)$ is an ergodic Markov process with unique invariant measure π .

- a) If X is strongly or exponentially ergodic then it is β -mixing.
 b) If X is uniformly ergodic, then it is strongly mixing.
 c) Any kind of exponential ergodic decay rate in L^2 imply the ρ -mixing.

Now we discuss ergodicity for the above examples of SDE (1) to derive mixing conditions. The exponential ergodicity of a Lévy driven OU process is established in Wang (2012). The result implies that if $\theta > 0$ then the α -Stable OU process (4) has a unique invariant measure π and is strongly ergodic (mixing). More generally, let's consider the case where the drift function f and diffusion function g satisfy sufficient conditions for the solution of (1) to exist and be unique (Bass et al. 2004; Fournier 2013). According to Kulik (2009), if $f(\cdot)$ is locally Lipschitz, and $\limsup_{|x| \rightarrow +\infty} \frac{f(x)}{x} < 0$, then, for any $\alpha \in (1, 2)$ and bounded function g , the solution of (1) is exponentially ergodic, and its invariant distribution exists and is unique. This result applies to the case of the stable OU model (4) under the condition $\theta > 0$.

In the case of stable driven CIR-models, some ergodicity conditions are well known when the driving process has only positive jumps. For $\theta > 0$ and $\mu \geq 0$, this model can be seen as a subcritical CBI process with an immigration rate μ . Thus, using the result of Li and Ma (2015), we conclude that the SCIR process is exponentially ergodic and hence strongly mixing. The exponential ergodicity is proved in Zhang et al. (2017). More precisely, it has been proven that if $\alpha \in (0, 2)$ and $\frac{\rho^2 c_\alpha}{2-\alpha}$, the SDE (7) with $q = 1$ is exponentially ergodic and hence strongly mixing. Recall that $c_\alpha = \frac{\alpha 2^{\alpha-1} \Gamma(\frac{\alpha+1}{2})}{\Gamma(1\frac{\alpha}{2})}$, where $\Gamma(\cdot)$ is the classical gamma function. For a more general SDE in the form (1), it was shown in Zhang and Zhang (2023), that the exponential ergodicity holds under some dissipative and non-degenerate assumptions on the drift f and diffusion function g . In terms of forthcoming study, we are currently interested in ergodicity properties for SDE in the form (1) under more different conditions as well as functional inequalities.

4.1 Numerical examples

We now scrutinize the numerical performance of our estimation procedure on synthetic data. Subsequently, we apply these procedures to real-world data. All the implemented codes were developed using the R software. For practical applications, we determine the optimal bandwidth $h = (h_n)$ using the method proposed by Sheather (2004). The R function `h.amize` facilitates this process. It is worth noting that, in many instances, people commonly opt for $h = h_n = n^{-1/5}$. A simulation study is conducted to evaluate the performance of the proposed estimation from data generated by ergodic (mixing) stationary processes.

4.1.1 A simulation studies

We simulate and approximate the solution $X = (X_t)_{t \geq 0}$ by using the Euler scheme on the interval $[0, T]$ with sample size $n = 500, 1000, 2000$ within a period $T = 1$. We

employ the exponential ergodicity results when choosing the model parameters. We consider the following parameters:

$$x_0 = 1, \quad \theta = 0.3, \quad \mu = 1.2 \quad \alpha = 1.6 \quad \beta = 0 \text{ or } 0.1 \quad \text{and} \quad \rho = 1.$$

Thus, Assumption (A_1) is verified for model (4) for general strictly driving α -stable process Z . For model (6), we consider positive jump driving stable process Z so that Assumption (A_1) is once again verified for $\beta > 0$. We choose $q = 2$ for model (5). For model (7) and according to Zhang et al. (2017), we choose (in order to have strongly mixing solution)

$$x_0 = 1, \quad \theta = 0.5, \quad \mu = 0.5 \quad \alpha = 1.5 \quad \beta = 0.1 \quad \text{and} \quad \rho = 0.5.$$

We generate data from models (4), (5) or (6) and (7) based on the previous configuration. To numerically validate the regularity condition on the density function of the stationary distribution π , assumed to be continuous in Assumption (A_2) , we plot the kernel density estimate of a realization of X alongside the histogram. Choosing a suitable time frequency (e.g., $\Delta_n = \frac{1}{\sqrt{n}}$), ensures that the models satisfy all the necessary conditions.

We assess the performance of the N-W estimator for the drift function and the estimated parameters of the diffusion part (α, β, ρ) in the context of the stable driven CIR, OU and Lotka-Volterra models on simulated data. Figures 4, 5 and 6 present the graphical performance of the N-W estimator concerning the true drift functions. The comparisons are made across three sample sizes $n = 500$, $n = 1000$ and $n = 2000$ with time interval $T = 1$ or $T = 10$. It is observed that varying T for a fixed n does not significantly improve the estimates of the drift function in the SCIR model. This confirms that the drift function cannot be precisely identified within a fixed time interval using the N-W estimator, regardless of how frequently the observations are sampled. Additionally, for a fixed length of the observation interval T , as the sample size increases, the N-W estimator does not exhibit better performance, aligning with the asymptotic theory of the N-W estimator for stochastic processes driven by Lévy motions. Figures 1 and 2 visually confirm the regularity condition on the density function of the stationary distribution. The performance of the estimation in Fig. 6 seems to imply that the model in (6) for $q = 2$ is ergodic. We have not found any references proving this. A current forthcoming work is the study of the ergodicity of SDE (1) under various conditions. To assess the performance, we employ not only visual illustrations through figures but also quantitative measures such as the Square Root of Average Square Errors, defined by:

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n \left(\hat{f}_n(x_i) - f(x_i) \right)^2,$$

where $f(x_i)$ and $\hat{f}_n(x_i)$ are transformed (Min-Max scaling) to cover the range of sample paths of X in a common scale, reducing the effect of outliers.

We summarized the performance in Tables 1, 4 and 5, which reports the results on the diffusion parameters (α, β, ρ) as well as the RMSE on n replicates with three sample sizes $n = 500$, $n = 1000$ and $n = 2000$, respectively for time interval $T = 1$ and $T = 10$. We can see that varying T for a fixed n slightly changes the estimates of the stable process parameters α, β , except for the scaling parameter ρ .

4.1.2 Real dataset

In this section, we shift our focus to the estimation using real data. We analyze financial data, specifically indexes or exchanges of the Canadian dollar against the US dollar, spanning a fixed period from May 2018 to June 2022. The analysis is conducted with a sample size of $n = 1000$ and $\Delta_n = \Delta = 1$.

Figures 3a–c present the time series variations of the exchange rates. To assess the stationarity of the time series, we examine the autocorrelation function (ACF) diagram (Alvarez and Olivares 2005). A rapidly decreasing ACF indicates stationarity, and the Dickey Fuller test confirms this with a p -value of 0.01. The ACF diagram demonstrates a quick decline, indicating weak dependence and implying the data's mixing properties, specifically ρ -mixing, which implies strong mixing. The density estimation of the stationary distribution is given also in Fig. 3a', b' and c' so that the regularity condition is satisfied.

In Figs. 7, we present the graphical N-W estimation of the observed drift function. To summarize the parameter estimation of the diffusion part, Tables 3 and 6 provide detailed results. Regarding the drift part, Tables 2 summarize the estimated drift coefficients based on real data. Additionally, in Fig. 8, we compare the prediction results with the stable OU model since the N-W estimator of the drift function seems to be linear and match better than the Stable CIR process.

Appendix: Tables and Figures

See Tables 1, 2, 3, 4, 5, 6 and Figs. 1, 2, 3, 4, 5, 6, 7, 8.

Table 1 Performance of the estimated diffusion and scaling parameters $(\hat{\alpha}, \hat{\beta}, \hat{\rho})$ with a standard symmetric 1.6-stable driven OU process

n and T	True parameters	Estimate parameters	RMSE
500 and T=1	$\alpha = 1.6, \beta = 0, \rho = 1$	$\hat{\alpha}_n = 1.70, \hat{\beta}_n = 0.08, \hat{\rho}_n = 0.80$	1.55
500 and T=10	$\alpha = 1.6, \beta = 0, \rho = 1$	$\hat{\alpha}_n = 1.69, \hat{\beta}_n = 0.015, \hat{\rho}_n = 3.38$	8.34
1000 and T=1	$\alpha = 1.6, \beta = 0, \rho = 1$	$\hat{\alpha}_n = 1.55, \hat{\beta}_n = 0.11, \hat{\rho}_n = 1.2$	1.10
1000 and T=10	$\alpha = 1.6, \beta = 0, \rho = 1$	$\hat{\alpha}_n = 1.56, \hat{\beta}_n = 0.13, \hat{\rho}_n = 4.87$	8.43
2000 and T=1	$\alpha = 1.6, \beta = 0, \rho = 1$	$\hat{\alpha}_n = 1.6346, \hat{\beta}_n = 0.1146, \hat{\rho}_n = 0.9335$	0.77
2000 and T=10	$\alpha = 1.6, \beta = 0, \rho = 1$	$\hat{\alpha}_n = 1.6029, \hat{\beta}_n = 0.064, \hat{\rho}_n = 4.33$	6.91

Table 2 Estimated drift coefficients $(\hat{\theta}, \hat{\mu})$ with the exchange rates data

n	Estimate parameters
1000	$\hat{\theta}_n = 1.08, \hat{\mu}_n = 0.93$
500	$\hat{\theta}_n = 1.06, \hat{\mu}_n = 0.95$
200	$\hat{\theta}_n = 1.028493, \hat{\mu}_n = 0.9674548$

Table 3 Estimated diffusion and scaling parameters $(\hat{\alpha}, \hat{\beta}, \hat{\rho})$ with the exchange rates data using an unknown strictly standard stable driven OU process

n	Estimated diffusion and scaling parameters
1000	$\hat{\alpha}_n = 1.98, \hat{\beta}_n = 1, \hat{\rho}_n = 13.21$
500	$\hat{\alpha}_n = 1.99, \hat{\beta}_n = 1, \hat{\rho}_n = 10.11$
200	$\hat{\alpha}_n = 1.99, \hat{\beta}_n = 1, \hat{\rho}_n = 6.17$

Table 4 Performance of the estimated diffusion and scaling parameters $(\hat{\alpha}, \hat{\beta}, \hat{\rho})$ with a standard strictly 1.5-stable Lotka Volterra process with positive jump activity $\beta > 0$ and $q = 1$

n and T	True parameters	Estimate parameters	RMSE
500 and T=1	$\alpha = 1.5, \beta = 0.1, \rho = 0.5$	$\hat{\alpha}_n = 1.51, \hat{\beta}_n = 0.38, \hat{\rho}_n = 0.51$	0.9699
1000 and T=1	$\alpha = 1.5, \beta = 0.1, \rho = 0.5$	$\hat{\alpha}_n = 1.59, \hat{\beta}_n = 0.12, \hat{\rho}_n = 0.43$	0.8033
2000 and T=1	$\alpha = 1.5, \beta = 0.1, \rho = 0.5$	$\hat{\alpha}_n = 1.54, \hat{\beta}_n = 0.11, \hat{\rho}_n = 0.45$	0.7980

Table 5 Performance of the estimated diffusion and scaling parameters ($\hat{\alpha}, \hat{\beta}, \hat{\rho}$) with a standard strictly 1.6-stable CIR process with positive jump activity $\beta > 0$ and $q = 2$

n and T	True parameters	Estimate parameters	RMSE
500 and T=1	$\alpha = 1.6, \beta = 0.1, \rho = 1$	$\hat{\alpha}_n = 1.66, \hat{\beta}_n = 0.13, \hat{\rho}_n = 0.88$	1.17
500 and T=10	$\alpha = 1.6, \beta = 0.1, \rho = 1$	$\hat{\alpha}_n = 1.51, \hat{\beta}_n = 0.08, \hat{\rho}_n = 5.95$	8.24
1000 and T=1	$\alpha = 1.6, \beta = 0.1, \rho = 1$	$\hat{\alpha}_n = 1.63, \hat{\beta}_n = 0.217, \hat{\rho}_n = 0.98$	3.72
1000 and T=10	$\alpha = 1.6, \beta = 0.1, \rho = 1$	$\hat{\alpha}_n = 1.53, \hat{\beta}_n = 0.135, \hat{\rho}_n = 6.32$	15.36
2000 and T=1	$\alpha = 1.6, \beta = 0.1, \rho = 1$	$\hat{\alpha}_n = 1.61, \hat{\beta}_n = 0.07, \hat{\rho}_n = 1.03$	2.07
2000 and T=10	$\alpha = 1.6, \beta = 0.1, \rho = 1$	$\hat{\alpha}_n = 1.64, \hat{\beta}_n = 0.04, \hat{\rho}_n = 4$	11.13

Table 6 Estimated diffusion and scaling parameters ($\hat{\alpha}, \hat{\beta}, \hat{\rho}$) with the exchange rates data using an unknown strictly standard SCIR process with given parameter q

n	q	Estimate parameters
1000	q=2	$\hat{\alpha}_n = 1.80, \hat{\beta}_n = -0.33, \hat{\rho}_n = 67.89$
1000	q=1.99	$\hat{\alpha}_n = 1.78, \hat{\beta}_n = -0.31, \hat{\rho}_n = 70.57$
1000	q=1.95	$\hat{\alpha}_n = 1.78, \hat{\beta}_n = -0.33, \hat{\rho}_n = 71.59$
1000	q=1.9	$\hat{\alpha}_n = 1.77, \hat{\beta}_n = -0.35, \hat{\rho}_n = 75.83$
1000	q=1.8	$\hat{\alpha}_n = 1.72, \hat{\beta}_n = -0.37, \hat{\rho}_n = 8736$
1000	q=1.6	$\hat{\alpha}_n = 1.59, \hat{\beta}_n = -0.28, \hat{\rho}_n = 130.70$
1000	q=1.5	$\hat{\alpha}_n = 1.53, \hat{\beta}_n = -0.20, \hat{\rho}_n = 163.02$
500	q=2	$\hat{\alpha}_n = 1.79, \hat{\beta}_n = -0.12, \hat{\rho}_n = 39.61$
500	q=1.99	$\hat{\alpha}_n = 1.78, \hat{\beta}_n = -0.12, \hat{\rho}_n = 39.99$
500	q=1.95	$\hat{\alpha}_n = 1.77, \hat{\beta}_n = -0.093, \hat{\rho}_n = 41.56$
500	q=1.9	$\hat{\alpha}_n = 1.75, \hat{\beta}_n = -0.06, \hat{\rho}_n = 43.74$
500	q=1.8	$\hat{\alpha}_n = 1.72, \hat{\beta}_n = -0.009, \hat{\rho}_n = 48.98$
500	q=1.6	$\hat{\alpha}_n = 1.59, \hat{\beta}_n = 0.08, \hat{\rho}_n = 70.13$
500	q=1.5	$\hat{\alpha}_n = 1.53, \hat{\beta}_n = 0.073, \hat{\rho}_n = 86.03$
200	q=2	$\hat{\alpha}_n = 1.69, \hat{\beta}_n = -0.31, \hat{\rho}_n = 25.56$
200	q=1.99	$\hat{\alpha}_n = 1.69, \hat{\beta}_n = -0.30, \hat{\rho}_n = 25.83$
200	q=1.95	$\hat{\alpha}_n = 1.67, \hat{\beta}_n = -0.25, \hat{\rho}_n = 26.98$
200	q=1.9	$\hat{\alpha}_n = 1.65, \hat{\beta}_n = -0.19, \hat{\rho}_n = 28.59$
200	q=1.8	$\hat{\alpha}_n = 1.60, \hat{\beta}_n = -0.008, \hat{\rho}_n = 32.42$
200	q=1.6	$\hat{\alpha}_n = 1.51, \hat{\beta}_n = 0.04, \hat{\rho}_n = 43.64$
200	q=1.5	$\hat{\alpha}_n = 1.44, \hat{\beta}_n = 0.08, \hat{\rho}_n = 5386$

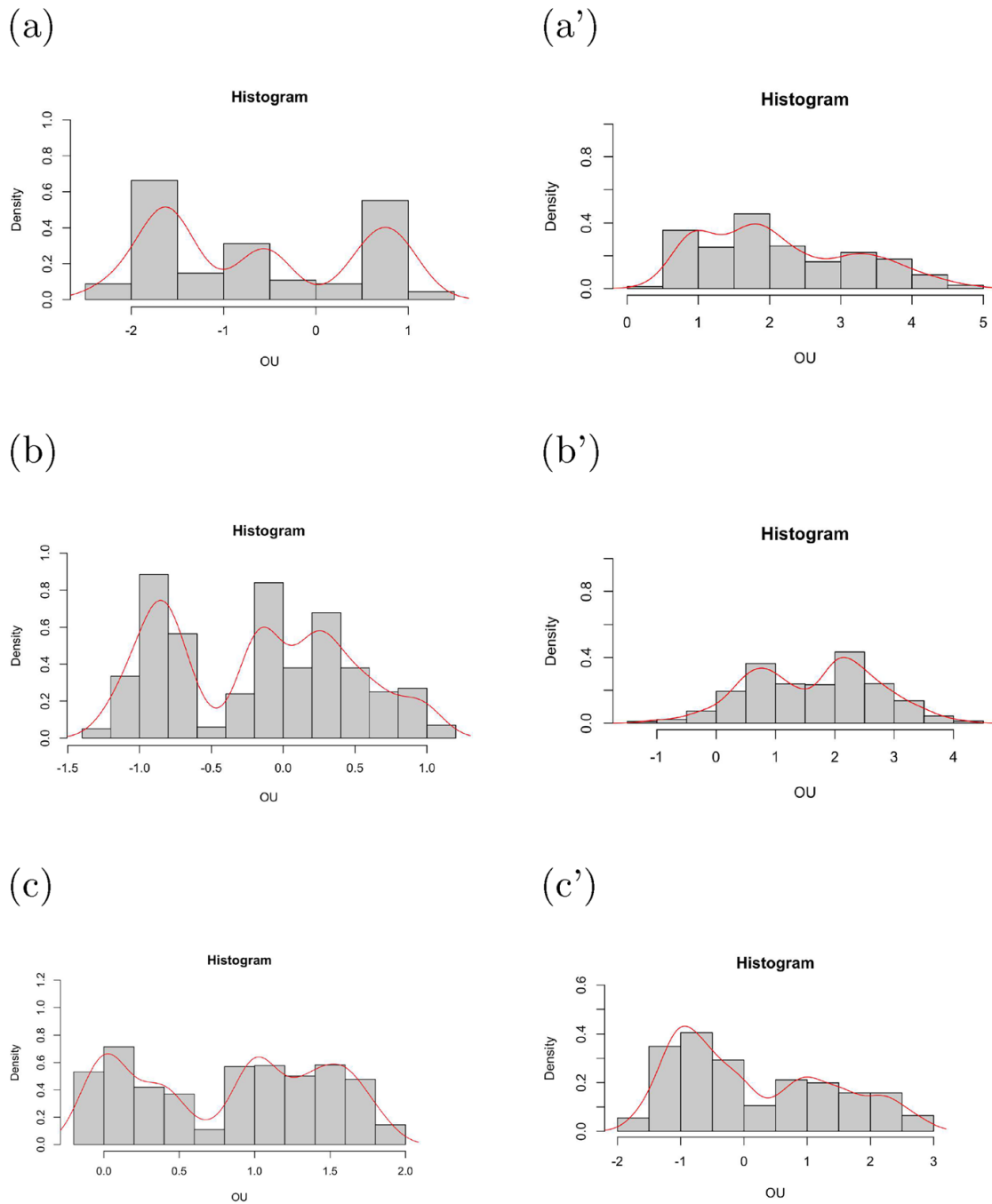


Fig. 1 Kernel density estimation of a standard symmetric 1.6-stable driven OU process with sample size $n = 500$ and $t = 1$ in (a), sample size $n = 500$ and $t = 10$ in (a'), sample size $n = 1000$ and $t = 1$ in (b), sample size $n = 1000$ and $t = 10$ in (b'), sample size $n = 2000$ and $t = 1$ in (c) and sample size $n = 2000$ and $t = 10$ in (c')

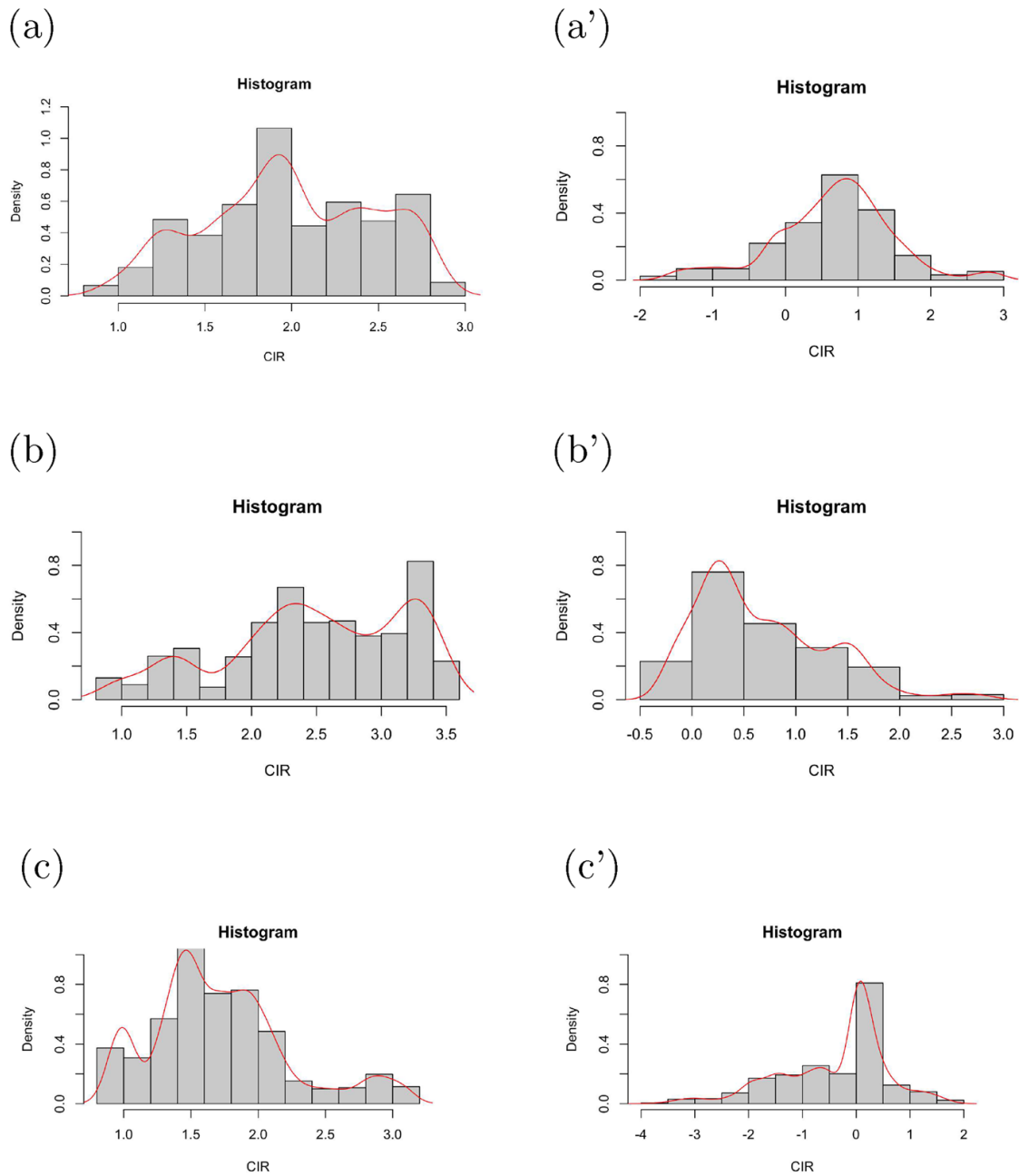


Fig. 2 Kernel density estimation of a standard strictly 1.6-stable CIR process with positive jump activity and sample size $n = 500$ and $t = 1$ in (a), sample size $n = 500$ and $t = 10$ in (a'), sample size $n = 1000$ and $t = 1$ in (b), sample size $n = 1000$ and $t = 10$ in (b'), sample size $n = 2000$ and $t = 1$ in (c) and sample size $n = 2000$ and $t = 10$ in (c')

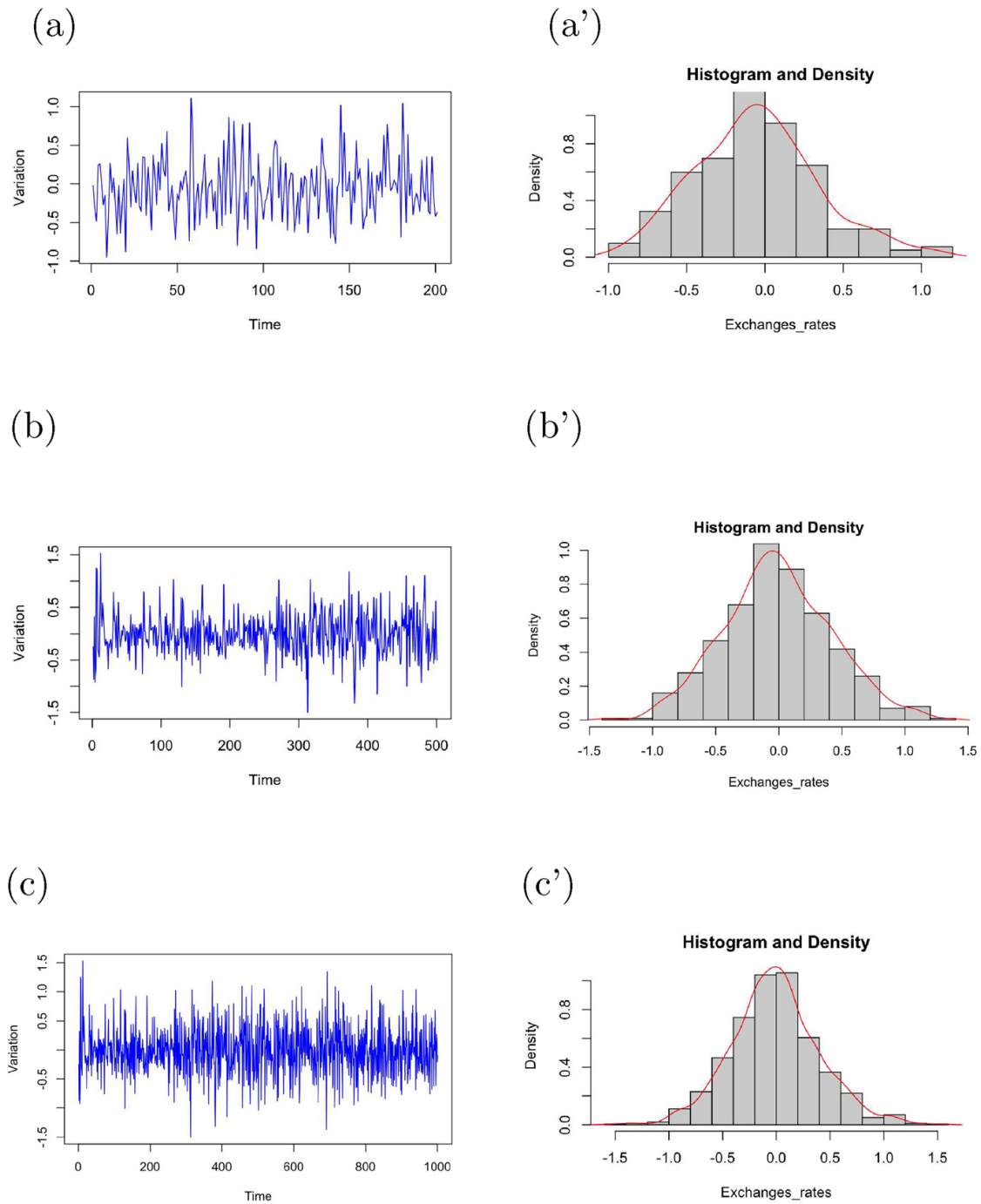


Fig. 3 The exchange rates of the Canadian dollar against the US dollar global variation and density estimation with sample size $n = 1000$ (a); local sample size $n = 500$ (b) and sample size $n = 200$ (c)

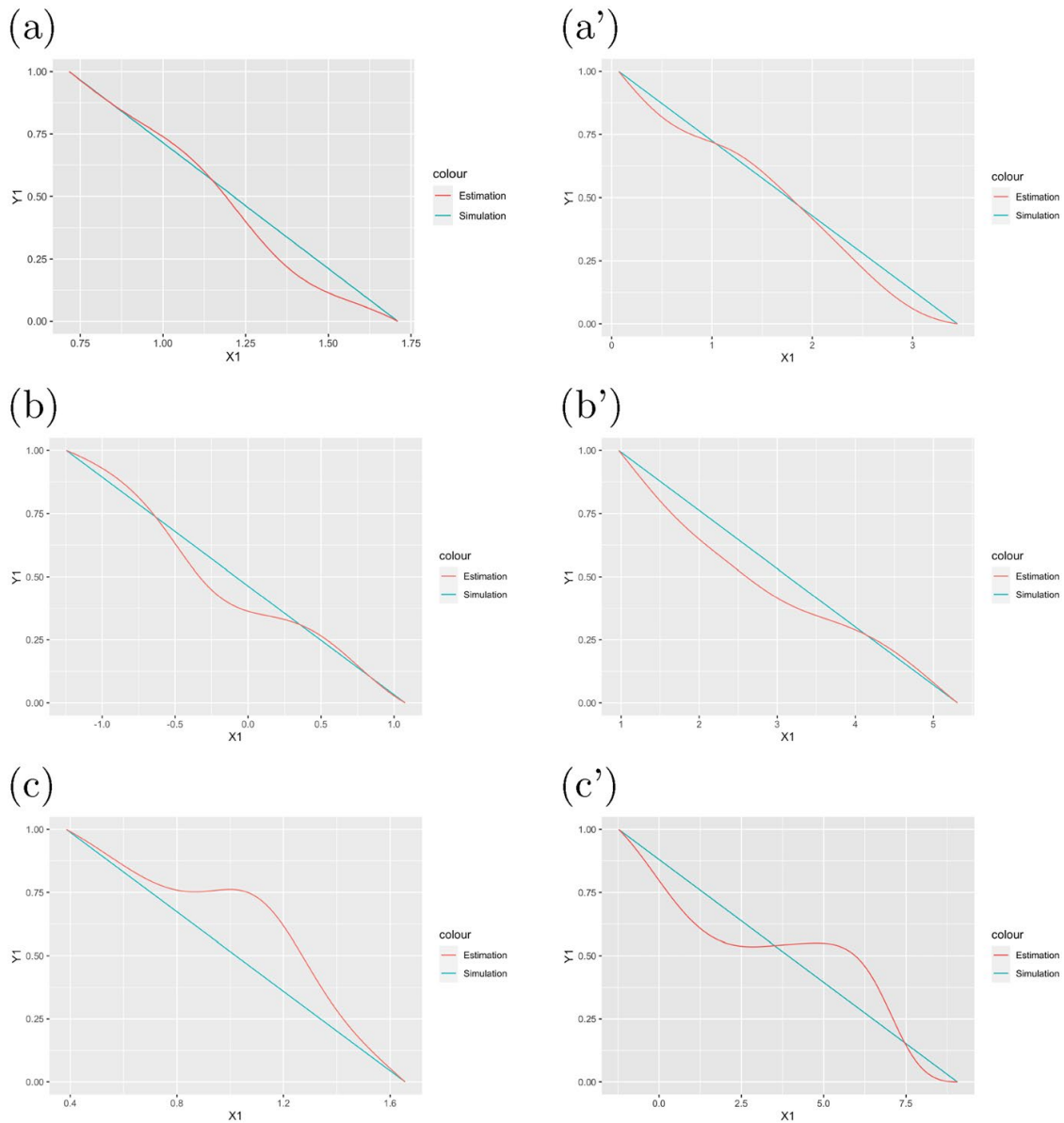


Fig. 4 Performance of the Nadaraya-Watson kernel estimator with respect to the true drift in the case of a standard symmetric 1.6-stable driven OU process with sample size $n = 500$ and $t = 1$ in **(a)**, sample size $n = 500$ and $t = 10$ in **(a')**, sample size $n = 1000$ and $t = 1$ in **(b)**, sample size $n = 1000$ and $t = 10$ in **(b')**, sample size $n = 2000$ and $t = 1$ in **(c)** and sample size $n = 2000$ and $t = 10$ in **(c')**

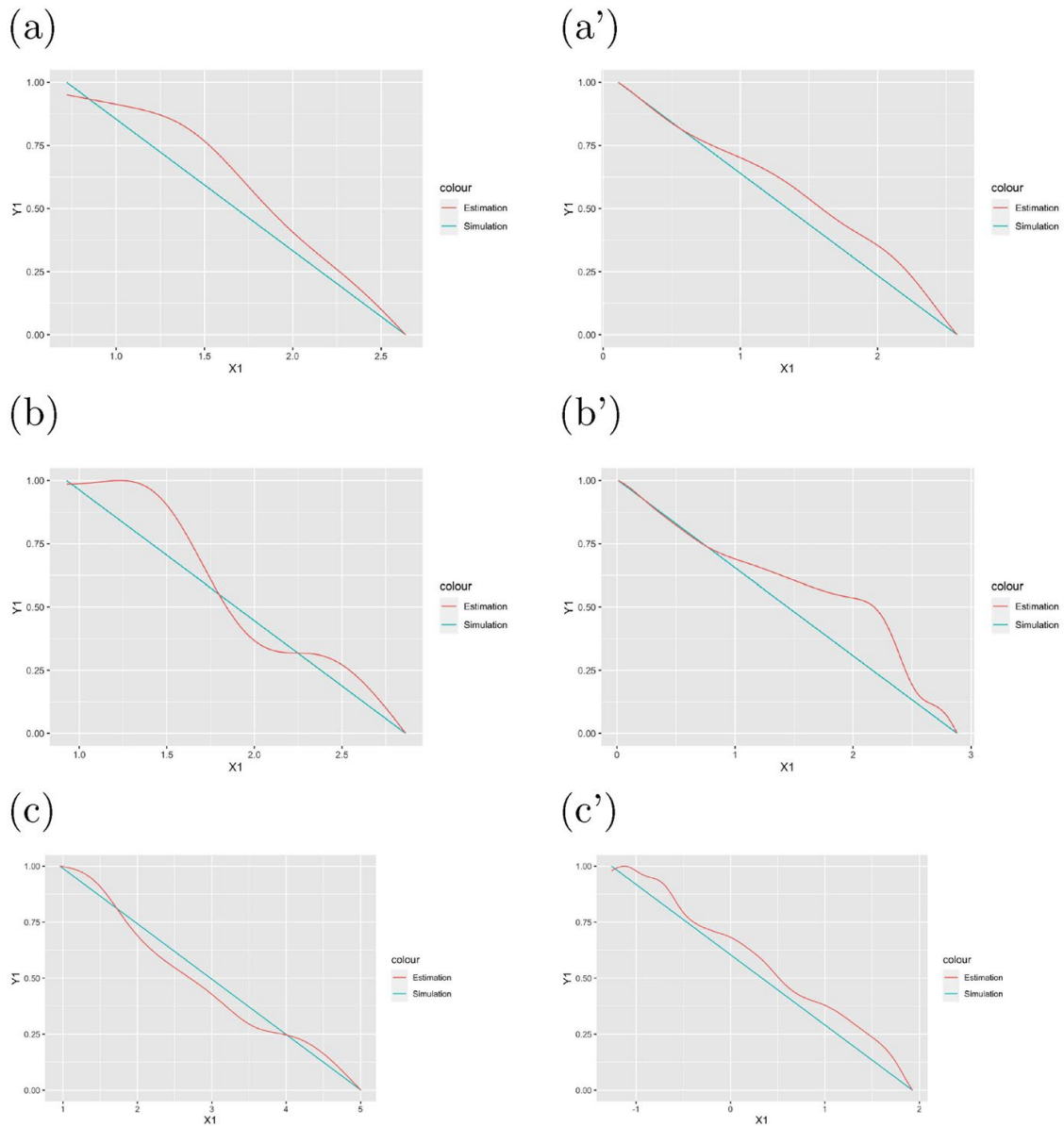


Fig. 5 Performance of the Nadaraya-Watson kernel estimator with respect to the true drift in the case of a standard strictly 1.6-stable CIR process with positive jump activity and sample size $n = 500$ and $t = 1$ in (a), sample size $n = 500$ and $t = 10$ in (a'), sample size $n = 1000$ and $t = 1$ in (b), sample size $n = 1000$ and $t = 10$ in (b'), sample size $n = 2000$ and $t = 1$ in (c) and sample size $n = 2000$ and $t = 10$ in (c')

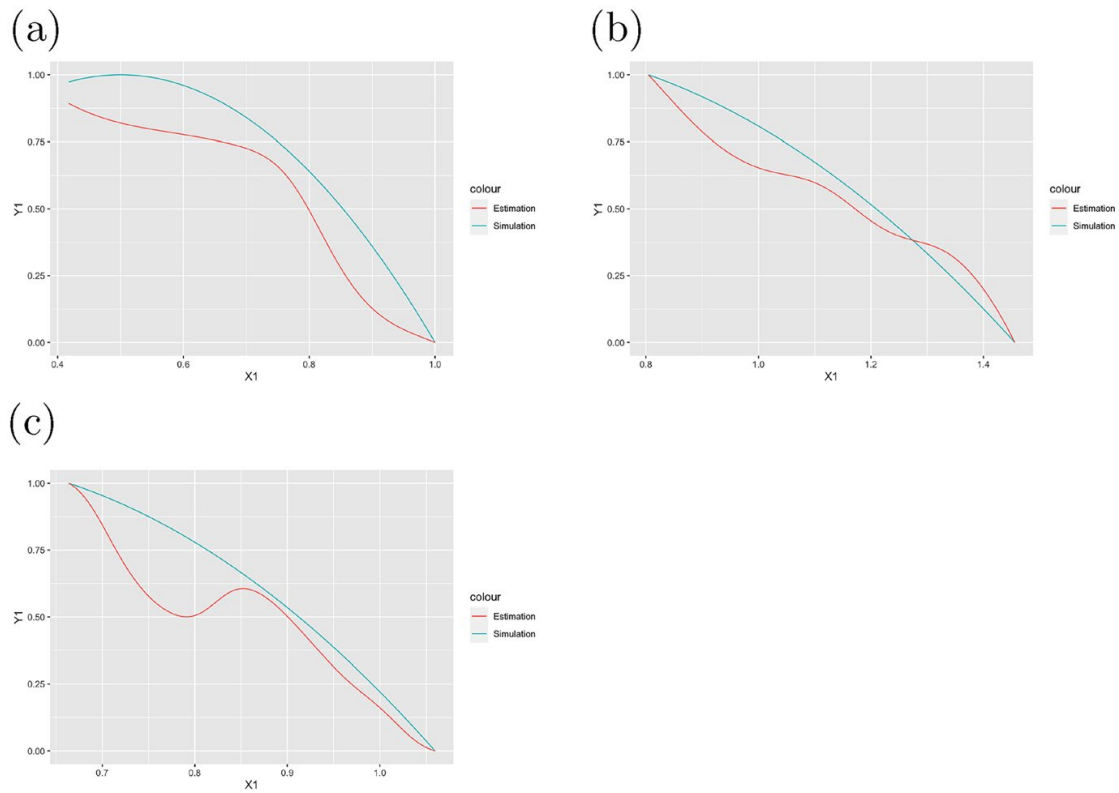


Fig. 6 Performance of the Nadaraya-Watson kernel estimator with respect to the true drift in the case of a standard and positive strictly 1.5-stable driven Lotka-Volterra process with sample size $n = 500$ and $t = 1$ in (a), sample size $n = 1000$ and $t = 1$ in (b), sample size $n = 2000$ and $t = 1$ in (c)

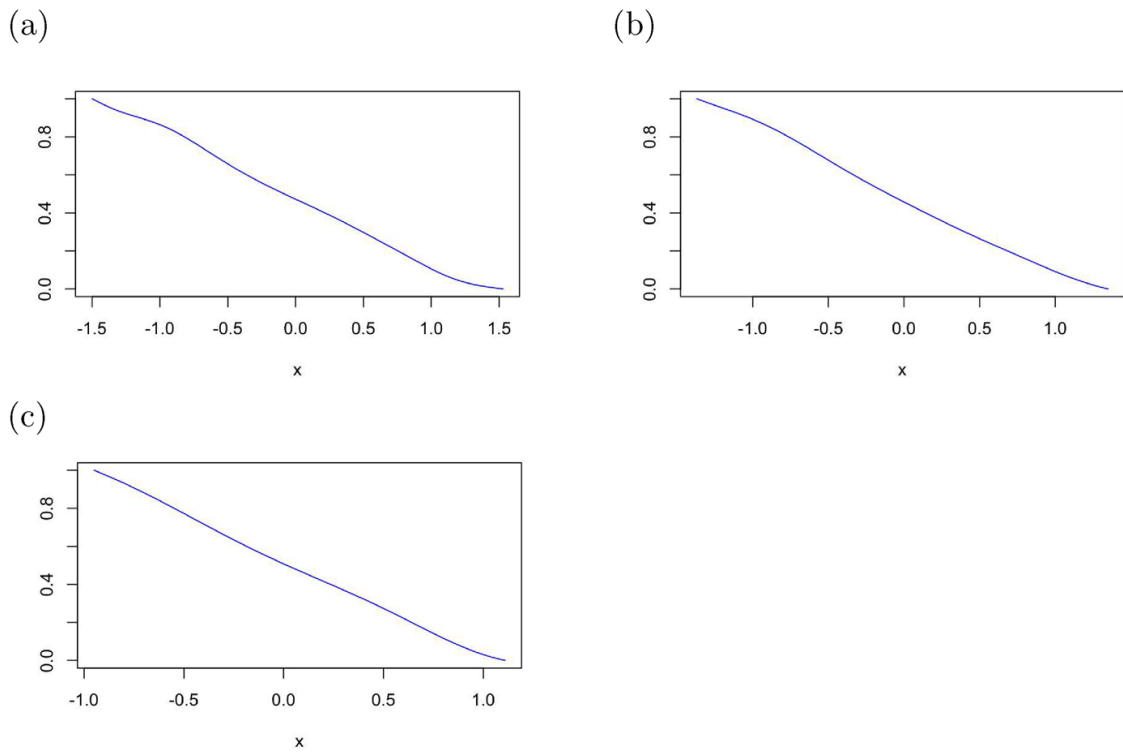


Fig. 7 Nadaraya-Watson kernel estimator of the drift function with the exchange rates of the Canadian dollar against the US dollar variation with the global sample size $n = 1000$ (a); local sample size $n = 500$ (b) and sample size $n = 200$ (c)

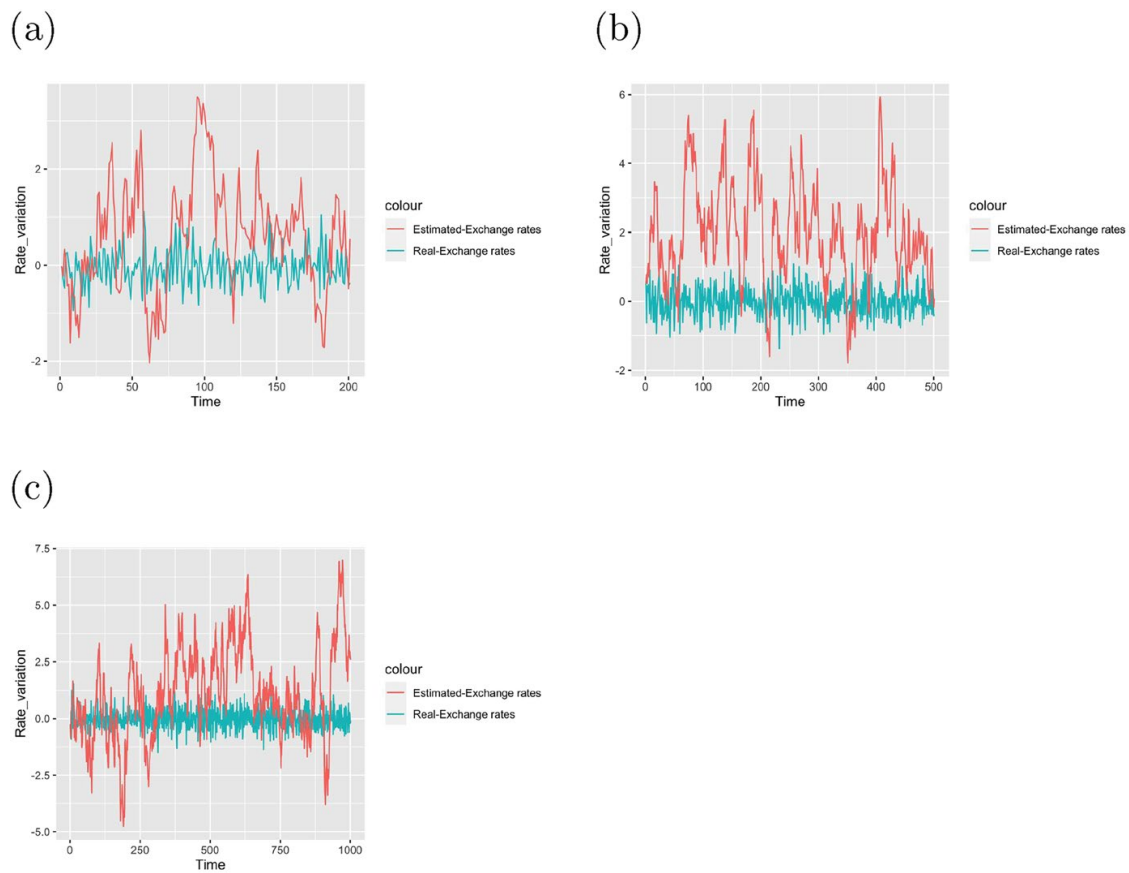


Fig. 8 Comparison between exchange rates prediction with Stable driven OU process with sample size $n = 200$ (a); sample size $n = 500$ (b) and sample size $n = 1000$ (c)

Acknowledgements Authors show their gratitude to the reviewers because their comments improved the paper.

Funding No funding received.

Data availability The real dataset used in this work is available online and free of charge. The R codes for the simulation study are available from the author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Consent for publication The author certify that the submission is an original work and is not under review at any other publication.

References

- Aït-Sahalia Y (2002) Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica* 70(1):223–262
- Allen LJS (2015) Stochastic population and epidemic models. *Math Biosci Lecture Stochas Biol Syst* 128

- Alvarez Alexander, Olivares Pablo (2005) Méthodes d'estimation pour des lois stables avec des applications en finance. *J de la société française de statistique* 146(4):23–54
- Applebaum D (2009) Lévy processes and stochastic calculus. Cambridge university press
- Bass RF, Burdzy K, Chen ZQ (2004) Stochastic differential equations driven by stable processes for which pathwise uniqueness fails. *Stochas Processes Appl* 111(1):1–15
- Bayraktar E, Clément E (2023) Estimation of a pure-jump stable cox-ingersoll-ross process. arXiv preprint [arXiv:2304.02386](https://arxiv.org/abs/2304.02386)
- Cattiaux P, Manou-Abi S (2014) Limit theorems for some functionals with heavy tails of a discrete time Markov chain. *ESAIM: Probability and Statistics*, 18:468–482
- Chambers JM, Mallows CL, BW4159820341 Stuck (1976) A method for simulating stable random variables. *J Am Stat Assoc* 71(354):340–344
- Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE (1996) On the lambert w function. *Adv Comput Math* 5:329–359
- Cox JC, Ingersoll JE Jr, Ross SA (2005) A theory of the term structure of interest rates. In *Theory of valuation*, pp 129–164. World Scientific
- Craigmile Peter, Herbei Radu, Liu Ge, Schneider Grant (2023) Statistical inference for stochastic differential equations. *Wiley Interdiscip Rev Comput Stat* 15(2):e1585
- Dexheimer N, Strauch C (2022) On lasso and slope drift estimators for Lévy-driven ornstein–uhlenbeck processes. arXiv preprint [arXiv:2205.07813](https://arxiv.org/abs/2205.07813)
- Dorogovcev AJ (1976) The consistency of an estimate of a parameter of a stochastic differential equation. *Theory Probab Math Stat* 10:73–82
- Fournier N (2013) On pathwise uniqueness for stochastic differential equations driven by stable lévy processes. *Annales de l'IHP Probabilités et statistiques* 49:138–159
- Zongfei F, Li Z (2010) Stochastic equations of non-negative processes with jumps. *Stochas Process Appl* 120(3):306–330
- Yaozhong H, Long H (2007) Parameter estimation for ornstein-uhlenbeck processes driven by stable lévy motions. *Commun Stochas Anal* 1(2):1
- Yaozhong H, Long H (2009) Least squares estimator for ornstein-uhlenbeck processes driven by α -stable motions. *Stochas Process Appl* 119(8):2465–2480
- Iacus SM et al. (2008) *Simulation and inference for stochastic differential equations: with R examples*, vol 486. Springer
- Janicki A, Michna Z, Weron A (1997) Approximation of stochastic differential equations driven by stable lévy motion. *Appl Math* 24(2):149–168
- Kasonga RA (1988) The consistency of a non-linear least squares estimator from diffusion processes. *Stochas Process Appl* 30(2):263–275
- Ken-Iti S (1999) Lévy processes and infinitely divisible distributions, vol 68. Cambridge university press
- Kogon SM, Williams DB (1998) Characteristic function based estimation of stable distribution parameters. *A practical guide to heavy tails: statistical techniques and applications*, pp 311–338
- Kulik AM (2009) Exponential ergodicity of the solutions to SDE's with a jump noise. *Stochas Process Appl* 119(2):602–632
- Kutoyants YA (2004) *Statistical inference for ergodic diffusion processes*. Springer Science & Business Media
- Breton AL (2009) On continuous and discrete sampling for parameter estimation in diffusion type processes. In: *Stochastic systems: modeling, identification and optimization, I*, pp 124–144. Springer
- Li PS, Li Z, Wang J, Zhou X (2022) Exponential ergodicity of branching processes with immigration and competition. arXiv preprint [arXiv:2205.15499](https://arxiv.org/abs/2205.15499)
- Li Z, Ma C (2015) Asymptotic properties of estimators in a stable cox-ingersoll-ross model. *Stochas Process Appl* 125(8):3196–3233
- Li Z, Mytnik L (2011) Strong solutions for stochastic differential equations with jumps. In: *Annales de l'IHP Probabilités et statistiques* 47, pp 1055–1067
- Lin ZY, Song YP, Yi JS (2014) Local linear estimator for stochastic differential equations driven by stable lévy motions. *Sci China Math* 57:609–626
- Lipcer RS, Liptser RS, Shirayev AN, Shirayev AN et al (2001) *Statistics of Random Processes II: II. Applications*, vol 2. Springer Science and Business Media
- Long H, Qian L (2013) Nadaraya-Watson estimator for stochastic processes driven by stable Lévy motions. *Electron J Stat*, 7, 1387–1418

- Maller RA, Müller G, Szimayer A (2009) Ornstein–uhlenbeck processes and extensions. Handbook of financial time series, pp 421–437
- Manou-Abi Solym (2023) Approximate solution for a class of stochastic differential equation driven by stable processes. Preprint, Submitted
- Manou-Abi SM (2015) Théorèmes limites et ordres stochastiques relatifs aux lois et processus stables. PhD thesis, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)
- Masuda H (2005) Simple estimators for non-linear markovian trend from sampled data: I. ergodic cases. MHF Preprint Series, 7
- Mikulevičius R, Fanhui X (2018) On the rate of convergence of strong Euler approximation for SDES driven by levy processes. Stochastics 90(4):569–604
- Nadaraya EA (1964) On estimating regression. Theory Probabil Appl 9(1):141–142
- Nolan JP (2020) Univariate stable distributions. Springer
- Pamen OM, Taguchi D (2017) Strong rate of convergence for the Euler-Maruyama approximation of SDES with hölder continuous drift coefficient. Stochas Process Appl 127(8):2542–2559
- Pardoux É (2016) Probabilistic models of population evolution: Scaling limits, genealogies and interactions, vol 1. Springer
- Rao BLSP (1983) Asymptotic theory for non-linear least squares estimator for diffusion processes. Statist J Theor Appl Stat 14(2):195–209
- Rao BLSP (2021) Nonparametric estimation of linear multiplier in stochastic differential equations driven by stable noise. arXiv e-prints, pp arXiv–2109
- Priola E (2012) Pathwise uniqueness for singular SDES driven by stable processes
- Privault N (2016) Stochastic calculus for jump processes. Unpublished working paper. Nanyang Technological University (<http://www.ntu.edu.sg/home/nprivault/index.html>)
- Rosinski J, Woyczynski WA (1986) On itô stochastic integration with respect to p-stable motion: inner clock, integrability of sample paths, double and multiple integrals. Annal Probabil, pp 271–286
- Samorodnitsky G, Taquq MS, Linde RW (1996) Stable non-gaussian random processes: stochastic models with infinite variance. Bull Lond Math Soc 28(134):554–555
- Sheather SJ (2004) Density estimation. Statist Sci. 588–597
- Shimizu Y (2006) M-estimation for discretely observed ergodic diffusion processes with infinitely many jumps. Stat Infer Stoch Process 9:179–225
- Shimizu Y, Yoshida N (2006) Estimation of parameters for diffusion processes with jumps from discrete observations. Stat Infer Stoch Process 9:227–277
- Taquq MS (1994) Stable non-Gaussian random processes: stochastic models with infinite variance. Chapman
- Wang J (2012) On the exponential ergodicity of lévy-driven ornstein-uhlenbeck processes. J Appl Probab 49(4):990–1004
- Watson GS (1964) Smooth regression analysis. Sankhyā Indian J Statist A 359–372
- Wei C (2020) Estimation for the discretely observed cox-ingersoll-ross model driven by small symmetrical stable noises. Symmetry 12(3):327
- Wu WB (2003) Nonparametric estimation for stationary processes. University of Chicago. Technic Rep 536
- Yang Xu (2017) Maximum likelihood type estimation for discretely observed CIR model with small α -stable noises. Statist Probabil Lett 120:18–27
- Zhang X, Zhang X (2023) Ergodicity of supercritical SDES driven by α -stable processes and heavy-tailed sampling. Bernoulli 29(3):1933–1958
- Zhang X, Yi H, Shu H (2019) Nonparametric estimation of the trend for stochastic differential equations driven by small α -stable noises. Stat Probabil Lett 151:8–16
- Zhang Z, Zhang X, Tong J (2017) Exponential ergodicity for population dynamics driven by α -stable processes. Stat Probabil Lett 125:149–159

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Solym M. Manou-Abi^{1,2,3} 

✉ Solym M. Manou-Abi
solym-mawaki.manou-
abi@umontpellier.fr; solym.manou-abi@mayotte.fr; solym.manou.abi@math.univ-poitiers.fr

¹ Institut Montpelliérain Alexander Grothendieck, UMR CNRS 5149, Université de Montpellier, Montpellier, France

² Département Sciences et Technologies, Université de Mayotte, Mayotte, France

³ Laboratoire de Mathématiques et Applications, UMR CNRS 7348, Université de Poitiers, Poitiers, France

Improved Estimators of Tail Index and Extreme Quantiles under Dependence Serials

Mamadou Aliou Barry¹, El Hadji Deme^{1*}, Aba Diop², and Solym M. Manou-Abi^{3,4}

¹*Laboratoire d'Etudes et de Recherche en Statistiques et Développement, Saint-Louis, Sénégal*

²*Département de Mathématiques, Université Alioune Diop de Bambey, Sénégal (Equipe de recherche en Statistique et Modèles Aléatoires (ERESMA)), Sénégal*

³*Institut Montpellierain Alexander Grothendieck, UMR CNRS 5149, Place Eugène Bataillon, 34090, Montpellier, France*

⁴*Centre Universitaire de Formation et de Recherche, Mayotte, France*

Received November 8, 2022; revised May 6, 2023; accepted June 5, 2023

Abstract—In this paper, we deal with the estimation problem for the extreme value parameters in the case of stationary β -mixing serials with heavy-tailed distributions. We first introduce two families of estimators generalizing the Hill's estimator. And from those families, three asymptotically unbiased estimators of the extreme value index are established. Our reflection is based on the generalized Jackknife methodology which consists of taking any pair of three special cases of our family of estimators to cancel the bias term. The resulting estimators are also used to deduce three asymptotically unbiased estimators of the extreme quantiles. In a simulation survey, the performance of our proposed methods are compared to alternative estimators recently introduced in the literature. Finally, our methods are applied to high financial losses data in order to estimate the Value-at-Risk of the daily stock returns on the S&P500 index.

DOI: 10.3103/S1066530723020011

Keywords: Estimation, Tail index, Extreme quantiles, Heavy-tailed, Bias reduction, Dependent serials

1. INTRODUCTION

Severe damage and losses often occur when unusual and intense events happen. For instance, extremely high water levels may cause a dike to smash and also large negative stocks return can produce a high financial losses or bankruptcy. Therefore, it is important to model these potential losses or their corresponding return periods in order to analyze and evaluate them. The Extreme Value Theory (EVT) helps to depict such events by estimating the parameter of rare events, i.e, tail risk.

The independence and identically distributed (i.i.d.) assumption is fundamental to classical EVT models. With such models, our main interest is on the estimation of the tail index parameter which is the basic parameter in extreme value theory. Several approaches for estimating this parameter have been proposed, including the famous Hill's estimator [14]. And its generalization has been investigated in this context [17].

However, some data such as financial or environmental data expose serial dependence, which can lead to biased estimates. To address this issue, some authors such as [4, 8], and [8] have proposed adapting existing estimators to the dependence case.

For instance, [4] proposed a method to adapt extreme value statistics to financial time series with serial dependence, by including additional conditions specifically to this type of data. In their work, they assumed that the time series is a β -mixing type, which allows them to take into account the dependence structure of the data. They showed that their method provides more accurate estimates of tail risk on

*E-mail: elhadji.deme@ugb.edu.sn

financial data compared to traditional estimators. We build upon the work of [4] by applying their method to estimate the tail parameter of a financial time series in the presence of serial dependence.

More precisely, in [4] Section 3, it is assumed that (X_1, X_2, \dots) is a β -mixing time series, that is, a series such that

$$\beta(m) := \sup_{p \geq 1} \mathbb{E} \left\{ \sup_{C \in \mathcal{B}_{p+m+1}^\infty} |\mathbb{P}(C|\mathcal{B}_1^p) - \mathbb{P}(C)| \right\} \rightarrow 0,$$

as $m \rightarrow \infty$, where \mathcal{B}_i^j denotes the σ -algebra generated by (X_i, \dots, X_j) . Without loss of generality, $\beta(m)$ measures the total variation distance between the unconditional distribution of the future of the time series and the conditional distribution of the future given the past of the time series when both are detached by m time points. Similarly to [4], Section 2, let F be the common marginal distribution function of $X_i, i \in \mathbb{N}$, assuming to be heavy-tailed (belonging to the the Fréchet domain of attraction), that is, there exist a positive number γ and the tail quantile function $U := (1/1 - F)^\leftarrow$ where $^\leftarrow$ denotes the left-continuous inverse function, such that

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^\gamma, \quad \forall x > 0. \tag{1.1}$$

From the relation in (1.1), U is said to be a regularly varying function at infinity with index γ . The parameter γ is called tail index or extreme value index and controls the behavior of the tail distribution function. Its estimation has mostly been studied in the case of i.i.d. random variables, although only few papers consider that for the case of time series with serial dependence features. We can cite among others, [7, 8] and [15] and very recently [4] and [3]. Furthermore, in the context of i.i.d. assumption, the simplest estimator for $\gamma > 0$ is the Hill's estimator [14] defined by

$$\hat{\gamma}_{k_n}^{(H)} := \frac{1}{k_n} \sum_{i=1}^{k_n} \log X_{n-i+1,n} - \log X_{n-k_n,n},$$

where $X_{1,n}, \dots, X_{n,n}$ stands for the order statistics and k_n represents an intermediate sequence, that is, a sequence such that $k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$, as $n \rightarrow \infty$. To prove the asymptotic normality of the tail index estimators such as the Hill's one, we need a second order condition which specifies the rate of convergence for the left-hand side in (1.1) to its limit. This condition can be formulated in different ways as shown below. We will use the formulation later-on.

Second order condition (C_{SO}). Suppose that there exists a positive or negative function A with $\lim_{t \rightarrow \infty} A(t) = 0$ and a real number $\rho < 0$ such that

$$\lim_{t \rightarrow \infty} \frac{1}{A(t)} \left(\frac{U(tx)}{U(t)} - x^\gamma \right) = x^\gamma \frac{x^\rho - 1}{\rho}, \quad \forall x > 0. \tag{1.2}$$

Note that in the case where $\rho = 0$, the right term in (1.2) equals to $x^\gamma \log x$. The rate of the convergence for the function A to 0 is essential since it helps to exhibit the bias term of the tail index estimators. From the assumption that the intermediate sequence k_n is such that $k_n^{1/2} A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, as $n \rightarrow \infty$ and assuming the following regularity conditions on the β -mixing coefficients (similar in [4], Section 3):

Regularity conditions (C_R). There exist $\epsilon > 0$, a bivariate function r and a sequence ℓ_n such that, as $n \rightarrow \infty$,

- (a) $\frac{\beta(\ell_n)}{\ell_n} n + \ell_n \frac{\log^2 k_n}{\sqrt{k_n}} \rightarrow 0$;
- (b) $\frac{n}{\ell_n k_n} Cov \left(\sum_{i=1}^{\ell_n} \mathbb{I}_{\{X_i > F^\leftarrow(1-k_n x/n)\}}, \sum_{i=1}^{\ell_n} \mathbb{I}_{\{X_i > F^\leftarrow(1-k_n y/n)\}} \right) \rightarrow r(x, y), \forall 0 \leq x, y \leq 1 + \epsilon$;
- (c) For some constant C : $\frac{n}{\ell_n k_n} \mathbb{E} \left[\left(\sum_{i=1}^{\ell_n} \mathbb{I}_{\{F^\leftarrow(1-k_n y/n) < X_i \leq F^\leftarrow(1-k_n x/n)\}} \right)^4 \right] \leq C(y - x), \forall 0 \leq x < y \leq 1 + \epsilon$ and $n \in \mathbb{N}$,

[8] established the asymptotic normality of $\hat{\gamma}_{k_n}^{(H)}$ as follows

$$\sqrt{k}(\hat{\gamma}_{k_n}^{(H)} - \gamma) \xrightarrow{d} \mathcal{N} \left(\frac{\lambda}{1 - \rho}, \sigma_H^2(\gamma) \right), \tag{1.3}$$

where $\sigma_H^2(\gamma) = \gamma^2 r(1, 1)$, with r the covariance structure, but has a simple expression in the i.i.d. context, where it is equal to γ^2 . In practice, the bias term of $\hat{\gamma}_{k_n}^{(H)}$ depends on whether ρ is close to zero or not, since under the second order condition (C_{SO}), the function $|A|$ is regularly varying at infinity with index ρ . This explains all the literature spread on bias correction in the i.i.d. context, see, e.g., [2, 10] and [16], etc. However, in the case of β -mixing time series, only two recent papers, [4] and [3], published in 2016 and 2018, respectively, deal with this problem and proposed a bias corrected estimator for γ . Moreover, they established their asymptotic properties under the regularity conditions (C_R) and the second order assumptions (C_{SO}).

This paper is organised as follows. In Section 3, we introduce the estimation and the asymptotic properties of three adaptive unbiased tail index estimators and their corresponding high quantiles. We recall that these estimators come from two families of estimators which are presented in [13] and generalize the Hill's estimator. And in Section 4, we consider examples of β -mixing series which are analogous to those in Section 5 of [4]. Then in Section 5, we match our theoretical results with a simulation assessment and real data analysis in order to highlight the salients of our methods. Finally, Section 6 is devoted to the proofs of our main results.

2. METHODOLOGY AND MAIN RESULTS

In [13], two families of tail index estimators are introduced in the case of i.i.d. of extreme values. These estimators are parameterized by a positive real α and they also generalize the Hill's estimator. Using a Jackknife approach on these estimators, meaning asymptotically unbiased estimators for γ , the tail index have been constructed. In this work, we fit in our methodology with the case of β -mixing sequences to estimate a tail parameter such as the extreme value index or an extreme quantile.

2.1. Adaptive Generalized Tail Index Estimators

We consider the statistics introduced in [5] and defined by:

$$M_{k_n}^{(\alpha)} := \frac{1}{k_n} \sum_{i=1}^{k_n} (\log X_{n-i+1,n} - \log X_{n-k_n,n})^\alpha, \quad \alpha > 0.$$

This can be rewritten as a functional of $(Q_n(t) := X_{n-[k_n t],n})_{t \in [0,1]}$, and the tail quantile process as follows (see [4], Appendix):

$$M_{k_n}^{(\alpha)} = \int_0^1 \left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha dt.$$

Now, to derive the asymptotic properties of this estimator, we first establish those of the quantile process and the moments $M_{k_n}^{(\alpha)}$. We show that the tail quantile process can be approximated by a Gaussian process as in the following results.

Proposition 2.1. *Suppose that (X_1, X_2, \dots) is a stationary β -mixing time series with continuous common marginal distribution function F and assume that (C_{SO}) and (C_R) hold. Let k_n be an intermediate sequence satisfying $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ and $k_n^{1/2} A(n/k_n) = O(1)$, as $n \rightarrow \infty$. Then, for given α , $\delta > 0$, under a Skorohod construction, there exist a function $\tilde{A} \sim A$, and a centered Gaussian process $(W(t))_{t \in [0,1]}$ with covariance function $r(\cdot, \cdot)$ defined in regularity conditions (C_R) , such that, as $n \rightarrow \infty$:*

$$\sup_{t \in (0,1)} t^{1/2+\delta} \left| \sqrt{k} \left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha - \gamma^\alpha (-\log t)^\alpha - \alpha \gamma^\alpha (-\log t)^{\alpha-1} (t^{-1}W(t) - W(1)) - \sqrt{k} \tilde{A}(n/k) \alpha \gamma^{\alpha-1} (-\log t)^{\alpha-1} \frac{t^{-\rho} - 1}{\rho} \right| \rightarrow 0 \text{ a.s.}$$

Note that the Proposition 2.1 holds when $\rho = 0$, in this case $(t^{-\rho} - 1)/\rho$ can be replaced by $-\log t$. But overall the the following, we focus on the case $\rho < 0$ because the case $\rho = 0$ necessitates a similar approach with different computation.

The aim of the following corollary is to provide the asymptotic properties of the moments $M_{k_n}^{(\alpha)}$ by applying the Proposition 2.1.

Corollary 2.1. *Assume that the conditions in the Proposition 2.1 hold. Then under the same Skorohod construction as in Proposition 2.1, as $n \rightarrow \infty$,*

$$\begin{aligned} & \sqrt{k_n} \left(M_{k_n}^{(\alpha)} - \gamma^\alpha \Gamma(\alpha + 1) \right) - \alpha \gamma^\alpha P^{(\alpha)} \\ & - \sqrt{k_n} \tilde{A}(n/k_n) \gamma^{\alpha-1} \frac{\Gamma(\alpha + 1)}{\rho} \left(\frac{1}{(1 - \rho)^\alpha} - 1 \right) \longrightarrow 0 \text{ a.s.}, \end{aligned}$$

where the process term

$$P^{(\alpha)} = \int_0^1 (-\log t)^{\alpha-1} (t^{-1}W(t) - W(1)) dt$$

is a normally distributed random variable with mean zero and covariance function $Cov(P^{(\alpha)}, P^{(\beta)}) = c_{\alpha,\beta}$ defined as,

$$c_{\alpha,\beta} = \int_0^1 \int_0^1 (-\log s)^{\alpha-1} (-\log t)^{\beta-1} \times \left(\frac{r(s,t)}{st} - \frac{r(s,1)}{s} - \frac{r(1,t)}{t} + r(1,1) \right) ds dt,$$

with the covariance structure r defined as in the regularity condition (C_R) .

The case $\alpha = 1$ is special since its corresponding tail index estimator $M_{k_n}^{(1)}$ is the Hill's estimator. For this reason, the asymptotic variance $\sigma_H^2(\gamma)$ is equal to $\gamma^2 c_{1,1}$, with $c_{1,1} = r(1,1)$ (see, eg. [8]). Following the idea in the Corollary 2.1 and the methodology in [13], we can define two families of tail index estimators which generalize the Hill's estimator. They are also parameterized by a positive real α . The first one is defined by:

$$\tilde{\gamma}_{k_n}^{(\alpha)} = \left(\frac{M_{k_n}^{(\alpha)}}{\Gamma(\alpha + 1)} \right)^{1/\alpha}, \quad \alpha > 0 \tag{2.4}$$

and the second one by:

$$\hat{\gamma}_{k_n}^{(\alpha)} = \frac{M_{k_n}^{(\alpha)}}{\Gamma(\alpha + 1) \left(M_{k_n}^{(1)} \right)^{\alpha-1}}, \quad \alpha \geq 1. \tag{2.5}$$

It is clear that the Hill's estimator $\hat{\gamma}_{k_n}^{(H)}$ corresponds to $M_{k_n}^{(1)} = \tilde{\gamma}_{k_n}^{(1)} = \hat{\gamma}_{k_n}^{(1)}$. The following theorem gives the asymptotic expansions of the estimators $\tilde{\gamma}_{k_n}^{(\alpha)}$ and $\hat{\gamma}_{k_n}^{(\alpha)}$ in terms of Gaussian process $P^{(\alpha)}$.

Theorem 2.1. *Assume that the conditions in the Proposition 2.1 hold. Then we have, as $n \rightarrow \infty$:*

$$\tilde{\gamma}_{k_n}^{(\alpha)} \stackrel{d}{=} \gamma + \frac{\gamma P^{(\alpha)}}{k_n^{1/2} \Gamma(\alpha + 1)} + \tilde{A}(n/k_n) \frac{1 - (1 - \rho)^\alpha}{\alpha \rho (1 - \rho)^\alpha} + o_{\mathbb{P}}(k_n^{-1/2}), \quad \text{for } \alpha > 0 \tag{2.6}$$

and for $\alpha \geq 1$

$$\hat{\gamma}_{k_n}^{(\alpha)} \stackrel{d}{=} \gamma + \frac{\gamma \alpha P^{(\alpha)}}{k_n^{1/2} \Gamma(\alpha + 1)} - \gamma(\alpha - 1) \frac{P^{(1)}}{k_n^{1/2}} + \tilde{A}(n/k_n) \left\{ \frac{1 - (1 - \rho)^\alpha}{\rho(1 - \rho)^\alpha} - \frac{\alpha - 1}{1 - \rho} \right\} + o_{\mathbb{P}}(k_n^{-1/2}). \tag{2.7}$$

The following corollary establishes the asymptotic normality of the generalized tail index estimators $\tilde{\gamma}_{k_n}^{(\alpha)}$ and $\hat{\gamma}_{k_n}^{(\alpha)}$.

Corollary 2.2. *Assume that the conditions in the Proposition 2.1 hold. Suppose that k_n is an intermediate sequence satisfying $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ and $\sqrt{k_n}A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, as $n \rightarrow \infty$. Then we have, as $n \rightarrow \infty$:*

$$\sqrt{k_n}(\hat{\gamma}_{k_n}^{(\alpha)} - \gamma) \xrightarrow{d} \mathcal{N} \left(\lambda \frac{1 - (1 - \rho)^\alpha}{\alpha \rho (1 - \rho)^\alpha}, \frac{\gamma^2 c_{\alpha, \alpha}}{\Gamma^2(\alpha + 1)} \right)$$

and

$$\xrightarrow{d} \mathcal{N} \left(\lambda \left\{ \frac{1 - (1 - \rho)^\alpha}{\rho (1 - \rho)^\alpha} - \frac{\alpha - 1}{1 - \rho} \right\}, \gamma^2 \left\{ (\alpha - 1)^2 c_{1,1} + \frac{c_{\alpha, \alpha}}{\Gamma^2(\alpha + 1)} - 2 \frac{(\alpha - 1) c_{1, \alpha}}{\Gamma(\alpha + 1)} \right\} \right),$$

where $c_{1,1}$, $c_{1,\alpha}$ and $c_{\alpha,\alpha}$ are as defined in the Corollary 2.1.

Remark 2.1. *From the Theorem 2.11, the results show that the bias terms in the i.i.d. case [13] remain unchanged in the case of β -mixing. This can be explained by the fact that the stationary β -mixing data satisfying the regularity condition (C_R) depend on the covariance function r which impact the variance terms of the families of estimators and not their bias terms. In addition, the analysis of [13] shows that the bias terms are minimal for some suitable values of (ρ, α) .*

2.2. Asymptotically Unbiased Estimators of the Tail Index

From the Corollary 2.2, it is clear that the second order parameter ρ is important for the reduction of the bias of the tail index estimators. In the case of i.i.d. random variables, the bias reduction problem have received much attention from many authors such as [1, 6, 11, 12, 22], and [27]. Nevertheless, only few authors considered the serial dependence topic. We can mention [7, 8] and [15], and more recently [4] and [3] as references.

The purpose in this section is to introduce a class of reduced bias estimators by using a jackknife approach. Following the idea of [12] and from the Remark 2.1, let's choose the values 1 and 2 for α . The resulting tail index estimators are:

$$\hat{\gamma}_{k_n}^{(H)} = \tilde{\gamma}_{k_n}^{(1)} = \hat{\gamma}_{k_n}^{(1)} = M_{k_n}^{(1)}, \quad \hat{\gamma}_{k_n}^{(2)} = \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \quad \text{and} \quad \tilde{\gamma}_{k_n}^{(2)} = \sqrt{\frac{M_{k_n}^{(2)}}{2}}. \tag{2.8}$$

From any pair of these above mentioned statistics, we construct the following adaptive reduced bias tail index estimators in the generalized Jackknife methodology sense (See eg, [12]):

$$\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub1)} := \frac{1}{\hat{\rho}} \left(\sqrt{2M_{k_n}^{(2)}} - (2 - \hat{\rho}) \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \right),$$

$$\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub2)} := \frac{1}{\hat{\rho}} \left((2 - \hat{\rho}) M_{k_n}^{(1)} - (1 - \hat{\rho}) \sqrt{2M_{k_n}^{(2)}} \right)$$

and

$$\hat{\gamma}_{k_n, \hat{\rho}}^{(dH)} := \frac{1}{\hat{\rho}} \left(M_{k_n}^{(1)} - (1 - \hat{\rho}) \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \right),$$

where $\hat{\rho}$ is either a canonical negative value $\hat{\rho} := \rho = \rho_0$ or a consistent estimator $\hat{\rho} := \hat{\rho}_{k_\rho}$ of ρ , with $k_\rho := k_{\rho, n}$ an intermediate sequence of integers greater than k_n , satisfying $k_\rho \rightarrow \infty$ and $k_\rho/n \rightarrow 0$, as $n \rightarrow \infty$. Note that the estimator $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$ can be written as in Eq. (4.2) of [4]

$$\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)} = M_{k_n}^{(1)} - \frac{M_{k_n}^{(2)} - 2 \left(M_{k_n}^{(1)} \right)^2}{2M_{k_n}^{(1)} \hat{\rho}_{k_\rho} (1 - \hat{\rho}_{k_\rho})^{-1}}.$$

Clearly, this estimator is seen as a corrected bias version of the Hill’s estimator $\hat{\gamma}_{k_n}^{(H)}$. It has been studied in the i.i.d. case by [22]. [18] showed that for the canonical choice $\rho_0 = -1$, the estimator $\hat{\gamma}_{k_n, \rho_0}^{(dH)}$ is the t-Hill log-gamma (t-lgHill) estimator of the tail index γ . They established its weak consistency for the moving average process. Recently, [4] adapted the estimator $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$ in the β -mixing serials case and established its asymptotic normality related to the β -mixing regularity conditions and a third order assumption.

The purpose of the following theorem is to establish only the asymptotic normality of the estimators $\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub1)}$, $\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub2)}$ and $\hat{\gamma}_{k_n, \hat{\rho}}^{(dH)}$ under the β -mixing regularity conditions and the second order assumption which is less strong than the third order condition (see (2.6) in [4]).

Theorem 2.2. *Let (X_1, X_2, \dots) be a stationary β -mixing time series with a continuous common marginal distribution function F and assume that (C_{SO}) and (C_R) hold. Let k_n be an intermediate sequence satisfying $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ and $\sqrt{k_n}A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, as $n \rightarrow \infty$. If $\hat{\rho}$ is either a canonical negative value $\hat{\rho} := \rho = \rho_0$ or an estimator $\hat{\rho} := \hat{\rho}_{k_\rho}$ of ρ , consistent in probability, with $k_\rho := k_{\rho, n}$ another intermediate sequence of integers greater than k_n and satisfying $k_\rho \rightarrow \infty$ and $k_\rho/n \rightarrow 0$, as $n \rightarrow \infty$, then we have:*

$$\sqrt{k_n} \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub1)} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\gamma, \rho) \right),$$

$$\sqrt{k_n} \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub2)} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\gamma, \rho) \right)$$

and

$$\sqrt{k_n} \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(dH)} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\gamma, \rho) \right),$$

where

$$\sigma^2(\gamma, \rho) = \frac{\gamma^2}{\rho} \left((2 - \rho)^2 c_{1,1} + (1 - \rho)^2 c_{2,2} + 2(2 - \rho)(\rho - 1)c_{1,2} \right).$$

Remark 2.2. Theorem 4.1 in [4] states a similar result for the estimator $\hat{\gamma}_{k_n}^{(dH)}(\hat{\rho}_{k_\rho})$. Comparing their theorem to the assumptions of our Theorem 2.2, we see that their conditions are less restrictive. In particular, we do not need a third order condition and only $\sqrt{k_\rho}A(n/k_\rho) \rightarrow \infty$, as $n \rightarrow \infty$ is needed on the intermediate sequence k_ρ . This is because we only require consistency for the estimator $\hat{\rho}_{k_\rho}$, not its asymptotic normality. However, our rate of convergence $\sqrt{k_n}$ is smaller than the one obtained in [4] since the assumption $\sqrt{k_n}A(n/k_n) \rightarrow \infty$, as $n \rightarrow \infty$, was used instead of our condition, $\sqrt{k_n}A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, as $n \rightarrow \infty$. Moreover, while their rate of convergence has the form $\sqrt{k_n}$, their intermediate sequence is larger than ours. This occurs when $|A|$, the rate function, varies regularly at infinity with index ρ .

Remark 2.3. As shown in Theorem 2.2, the three estimators are asymptotically equivalent. This is advantageous because it provides flexibility in choosing which estimator to use in practice, while still obtaining similar results. This can be particularly useful in situations where one estimator may be better suited for a particular purpose or when computational efficiency is a concern.

However, it is important to note that although the estimators may be asymptotically equivalent, they may still differ in terms of their finite-sample properties, such as bias or efficiency. Therefore, it is always a good idea to consider these factors when choosing an estimator in practice, and to ensure that the estimator is appropriate for the specific problem at hand.

A possible choice for $\hat{\rho}_{k_\rho}$ is the most performed estimator among those studied in the i.i.d. case (see, e.g, [6, 12]) and used in the β -mixing case by [3, 4]:

$$\hat{\rho}_{k_\rho}^{(*)} = \frac{6S_{k_\rho}^{(2)} - 4 + \sqrt{3S_{k_\rho}^{(2)} - 2}}{4S_{k_\rho}^{(2)} - 3}, \quad \text{provided } S_{k_\rho}^{(2)} \in \left(\frac{2}{3}, \frac{3}{4} \right), \tag{2.9}$$

where

$$S_{k_\rho}^{(2)} = \frac{3}{4} \frac{\left[M_{k_\rho}^{(4)} - 24 \left(M_{k_\rho}^{(1)} \right)^4 \right] \left[M_{k_\rho}^{(2)} - 2 \left(M_{k_\rho}^{(1)} \right)^2 \right]}{\left[M_{k_\rho}^{(3)} - 6 \left(M_{k_\rho}^{(1)} \right)^3 \right]^2}.$$

The asymptotic properties of $\hat{\rho}_{k_\rho}^{(*)}$ have been established in [4] in the case of β -mixing serials.

Since the mentioned jackknife estimators have similar asymptotic variances, they can be denoted by $\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet)}$. From the previous remark, follow the direct consequences in the corollary:

Corollary 2.3. *Let (X_1, X_2, \dots) be a stationary β -mixing time series with a continuous common marginal distribution function F and assume that (C_{SO}) and (C_R) hold. Let k_n be an intermediate sequence satisfying $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ and $\sqrt{k_n}A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, as $n \rightarrow \infty$. If $\hat{\rho}_{k_\rho}^{(*)}$ is the estimator of ρ defined in (2,9), where the intermediate sequence $k_\rho := k_{\rho, n}$ is greater than k_n and satisfies $k_\rho \rightarrow \infty$, $k_\rho/n \rightarrow 0$ and $\sqrt{k_\rho}A(n/k_\rho) \rightarrow \infty$, as $n \rightarrow \infty$, then we have:*

$$\sqrt{k_n}(\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}^{(*)}}^{(\bullet)} - \gamma) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\gamma, \rho)),$$

where $\sigma^2(\gamma, \rho)$ is given in Theorem 2.2.

2.3. Estimation of Extreme Quantiles

The quantile, at probability level $(1 - t) \in (0, 1)$ with respect to F denoted by $x(t)$, is defined as:

$$x(t) := U(1/t). \tag{2.10}$$

Therefore $x(t)$, the quantile is estimated by $\hat{x}(t) := X_{n-[nt];n}$. Consider some positive sequence $t = t_n$ which tends to 0, as $n \rightarrow \infty$. Then it is possible to establish the consistency of $\hat{x}(t_n)$, the non parametric estimator when $t_n \rightarrow 0$ slowly enough. However, in some fields of life such as insurance, finance, hydrology and reliability, a major requirement is to find values, large enough so that the chances of exceeding them are very small. This leads to removing the restriction on the rate of convergence of t_n to 0. Moreover, the interest is to estimate $x(p)$, an extreme quantile, where p , the tail probability, depends on the observed sample size n (i.e $p := p_n$) and p_n is smaller than $1/n$. Hence, it is not possible to have a non parametric estimate of such a quantile.

The goal of this section is to tackle this estimation problem in a β -mixing serials framework in order to estimate $x(p) = U(1/p)$, the extreme quantile with $np < 1$. Following [4, 20], we propose, in a biased reduction method, to estimate $x(p)$, the extreme quantile. The construction of our bias reduction procedure is based on our second-order condition (C_{SO}) which states

$$\frac{U(tx)}{U(t)} \approx x^\gamma \left\{ 1 - \rho^{-1} A(t)[1 - x^\rho] \right\}, \quad t \rightarrow \infty.$$

Let $tx = 1/p$ and $t = n/k_n \rightarrow \infty$, as $n \rightarrow \infty$. We obtain the following approximation:

$$x(p) = U(1/p) \approx U(n/k_n) \left(\frac{k_n}{np} \right)^\gamma \left\{ 1 - \rho^{-1} A(n/k_n) \left[1 - \left(\frac{k_n}{np} \right)^\rho \right] \right\}, \tag{2.11}$$

where γ , ρ and $A(n/k_n)$ are unknown. The first part $U(n/k_n) (k_n/(np))^\gamma$ in the right side of (2.11) is exactly estimated by the Weissman's estimator [26] $\hat{x}_{k_n}^{(W)}(p)$ and defined as:

$$\hat{x}_{k_n}^{(W)}(p) = X_{n-k_n, n} \left(\frac{k_n}{np} \right)^{\hat{\gamma}_{k_n}^{(H)}},$$

where $X_{n-k_n, n}$ is the empirical estimator of $U(n/k_n)$ and $\hat{\gamma}_{k_n}^{(H)}$ is the Hill's estimator of γ . Obviously, $\hat{x}_{k_n}^{(W)}(p)$, the Weissman's estimator, exhibits a potential bias because it depends on the Hill's estimator

Table 1. Estimated values of the Value at Risk with 95% confidence intervals at the optimal point of the S&P500 negative log-return series. These values are computed with $\hat{\rho}_{k_\rho} = -1.468$ which corresponds to the value of $k_\rho = 442$

Tail index			Value at Risk						
$\hat{\gamma}_{k_n}^{(\bullet)}$	k_n^*	$\hat{\gamma}_{k_n^*}^{(\bullet)}$	$\hat{x}_{k_n^*}^{(\bullet)}(p)$	$p = 0.01$	95%-Conf. Int	Cover	$p = 0.001$	95%-Conf. Int	Cover
$\hat{\gamma}_{k_n}^{(H)}$	68	0.4665	$\hat{x}_{k_n^*}^{(W)}(p)$	0.0597	(0.0575, 0.1418)	0.0843	0.1752	(0.1602, 0.5949)	0.4347
$\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}$	104	0.4005	$\hat{x}_{k_n^*, \hat{\rho}_{k_\rho}}^{(Ub1)}(p)$	0.0552	(0.0545, 0.1017)	0.0472	0.1393	(0.1322; 0.5487)	0.4165
$\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$	126	0.3919	$\hat{x}_{k_n^*, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$	0.0511	(0.0129, 0.0594)	0.0465	0.1264	(0.0476, 0.1879)	0.1403
$\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$	116	0.4006	$\hat{x}_{k_n^*, \hat{\rho}_{k_\rho}}^{(dH)}(p)$	0.0543	(0.0538, 0.0710)	0.0172	0.1369	(0.1302, 0.4352)	0.3050

$\hat{\gamma}_{k_n}^{(H)}$ which has a similar problem. The expression $1 - \rho^{-1}A(n/k_n)[1 - (k_n/(np))^p]$ can be viewed as a correcting term since $A(n/k_n)$ tends to 0. And this leads to the need to estimate $A(n/k_n)$ and ρ .

Using the tail index estimators in (2.8) and their expansions in Theorem 2.1, we obtain

$$\frac{\frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} - M_{k_n}^{(1)}}{A(n/k_n)} \xrightarrow{\mathbb{P}} \frac{\rho}{(1 - \rho)^2}. \quad \text{As } n \rightarrow \infty \tag{2.12}$$

From equation (2.12), by substituting ρ with $\hat{\rho}$, we estimate $A(n/k_n)$ as

$$\hat{A}(n/k_n) = \frac{(1 - \hat{\rho})^2}{\hat{\rho}} \left(\frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} - M_{k_n}^{(1)} \right).$$

Finally, the use of the estimators and the approximation in Section 2.2 and (2.11) respectively lead to the following class of reduced bias estimators of $x(p)$, the extreme quantile:

$$\begin{pmatrix} \hat{x}_{k_n, \hat{\rho}}^{(Ub1)}(p) \\ \hat{x}_{k_n, \hat{\rho}}^{(Ub2)}(p) \\ \hat{x}_{k_n, \hat{\rho}}^{(dH)}(p) \end{pmatrix} = X_{n-k_n, n} \times \begin{pmatrix} \left(\frac{k}{np}\right)^{\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub1)}} \\ \left(\frac{k}{np}\right)^{\hat{\gamma}_{k_n, \hat{\rho}}^{(Ub2)}} \\ \left(\frac{k}{np}\right)^{\hat{\gamma}_{k_n, \hat{\rho}}^{(dH)}} \end{pmatrix} \times \left\{ 1 - \frac{M_{k_n}^{(2)} - M_{k_n}^{(1)}}{\hat{\rho}^2 (1 - \hat{\rho})^{-2}} \left[1 - \left(\frac{k}{np}\right)^{\hat{\rho}_{k_\rho}} \right] \right\}. \tag{2.13}$$

The asymptotic normality of extreme quantile estimators is established in the following theorem.

Theorem 2.3. *Let (X_1, X_2, \dots) be a stationary β -mixing time series with a continuous common marginal distribution function F and assume that (C_{SO}) and (C_R) hold. Let $\hat{\rho}$ be either a canonical negative value $\hat{\rho} := \rho = \rho_0$ or an estimator $\hat{\rho} := \hat{\rho}_{k_\rho}$ of ρ , consistent in probability such that $k_\rho := k_{\rho, n}$ is an intermediate sequence of integers satisfying $k_\rho \rightarrow \infty$ and $k_\rho/n \rightarrow 0$, as $n \rightarrow \infty$. Consider another intermediate sequence k_n smaller than $k_{\rho, n}$, such that $k_n \rightarrow \infty$, $k_n^{1/2}A(n/k_n) \rightarrow \lambda \in \mathbb{R}$, $np_n/k_n \rightarrow 0$ and $\log(np_n)/\sqrt{k_n} \rightarrow 0$. Then*

$$\frac{\sqrt{k_n}}{\log(np_n)} \left(\frac{\hat{x}_{k_n, \hat{\rho}}^{(\bullet)}(p)}{x(p)} - 1 \right) \rightarrow \mathcal{N}(0, \sigma^2(\gamma, \rho)),$$

where $\sigma^2(\gamma, \rho)$ is defined in Theorem 2.2.

3. EXAMPLES OF β -MIXING SERIES

In this section, we present the examples dealing with classical stationary models satisfying the regularity (C_R) assumptions (see [4], Section 5).

Example 3.1 (Autoregressive (AR) model). Consider the stationary solution of AR(1) equation:

$$X_i = \theta X_{i-1} + \varepsilon_i, \tag{3.14}$$

for some $\theta \in (0, 1)$ and i.i.d. random variables ε_i . The distribution function of the innovations is denoted by F_ε . Assume that F_ε admits a positive Lebesgue density which is L_1 -Lipschitz-continuous; see [7] Eq. (42). Suppose that as $x \rightarrow \infty$,

$1 - F_\varepsilon(x) \sim px^{-1/\gamma}l(x)$ and $F_\varepsilon(-x) \sim qx^{-1/\gamma}l(x)$, for some slowly varying function l and $p = 1 - q \in (0, 1)$. Then from Sect. 3.2 of [7], we obtain $1 - F(x) \sim d_\theta(1 - F(x))$ as $x \rightarrow \infty$, where $d_\theta = (1 - \theta^{1/\gamma})^{-1}$. Furthermore, the regularity conditions hold with,

$$r(x, y) = x \wedge y + \sum_{m=1}^{\infty} (c_m(x, y) + c_m(y, x));$$

where $c_m(x, y) = x \wedge y \theta^{m/\gamma}$.

Example 3.2 (Moving average (MA) model). Consider the stationary solution of MA(1) equation:

$$X_i = \theta \varepsilon_{i-1} + \varepsilon_i, \tag{3.15}$$

where the innovation ε satisfies the same conditions as in the AR(1) model (Example 3.1). And from Sect. 3.2 of [7], we obtain $1 - F(x) \sim d_\theta(1 - F_\varepsilon(x))$ as $x \rightarrow \infty$, where $d_\theta = 1 + \theta^{1/\gamma}$. One can also compute $r(x, y) = x \wedge y + (1 + \theta^{1/\gamma})^{-1}(x \wedge y \theta^{1/\gamma} + y \wedge x \theta^{1/\gamma})$.

Example 3.3 (Generalized autoregressive conditional heteroskedasticity (GARCH) model). Consider the stationary solution to the recursive system of equations:

$$\begin{cases} X_t = \sigma_t \varepsilon_t, \\ \sigma_t^2 = \lambda_0 + \lambda_1 X_{t-1}^2 + \lambda_2 \sigma_{t-1}^2, \end{cases} \tag{3.16}$$

where the process of innovations ε are i.i.d. with zero mean and unit variance. The stationary solution X_t of this GARCH(1,1) model follows a heavy-tailed distribution, even if the innovations ε_t are normally distributed, see [19] and [9].

This GARCH(1,1) model satisfies the regularity conditions (C_R) (see [24] and [8]), but without an explicit covariance function $r(., .)$.

4. SIMULATION STUDY

In this section, we proceed by generating the data for the four models (similarly in [4], Section 6). This involves an independent model and the three models mentioned above, Examples 3.1–3.3. We define ε such that:

$$F_\varepsilon(u) = \begin{cases} (1 - q)(1 - \tilde{F}(-u)) & \text{if } u < 0, \\ 1 - q + q\tilde{F}(u) & \text{if } u > 0, \end{cases}$$

where \tilde{F} stands for the unit Fréchet distribution function and for $q = 0.75$. F_ε belongs to the max-domain of attraction with extreme index $\gamma = 1$. We generate $N = 1000$ time series of size $n = 1000$ based on i.i.d. observations from F_ε . This concerns all four models. The theoretical value of $\gamma = 1$ concerns the first three models. For the fourth model $\gamma = 0.258$:

- **Model 1.** Independence model with $X_t = \varepsilon_t$ (as a particular case of AR(1) or MA(1) with $\theta = 0$). The theoretical value of $x(0.001)$ is 749.80.
- **Model 2.** AR(1) model (3.14) with $\theta = 0.3$. The theoretical value of $x(0.001)$ is 1072.26.
- **Model 3.** MA(1) model (3.15) with $\theta = 0.3$. The theoretical value of $x(0.001)$ is 972.85.

- **Model 4.** *GARCH*(1, 1) model (5.17) with standardized Student t innovations with 5.64 degrees of freedom and $\lambda_0 = 8.26 \times 10^{-07}$, $\lambda_1 = 0.052$, $\lambda_2 = 0.941$. The theoretical value of $x(0.001)$ is 0.0592.

Note that the theoretical values were computed by the Monte Carlo method based on 500 samples of size 10^6 , as in pg 334 of [4]. In our simulation study, we generate $N = 1000$ samples of size $n = 1000$. Firstly, we focus on the extreme value index γ . We apply the four tail index estimators: $\hat{\gamma}_{k_n}^{(H)}$, $\hat{\gamma}_{k_n, k_\rho}^{(Ub1)}$, $\hat{\gamma}_{k_n, k_\rho}^{(Ub2)}$ and $\hat{\gamma}_{k_n, k_\rho}^{(dH)}$, with their associated extreme quantile estimators $\hat{x}_{k_n}^{(W)}(p)$, $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}(p)$, $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$ and $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}(p)$, with $p = 0.001$, on each model in order to compare them. But for the estimator $\hat{\rho}_{k_\rho}$, we use $\hat{\rho}_{k_\rho} = \hat{\rho}_{k_\rho}^{(*)}$ defined in (2,9), where the sequence k_ρ is selected as follows (similarly to [4], Section 6.2):

$$k_\rho := \sup \left\{ k : k \leq \min \left(m - 1, \frac{2m}{\log \log m} \right) \text{ and } \hat{\rho}_k \text{ exists} \right\},$$

where m is the number of positive observations in the sample.

Secondly, on the one hand we compare the performance of the tail index estimators in each model. And on the other, we do the same for the extreme quantiles estimators. For this reason, we compute the absolute value of the mean of the bias (ABias) together with the root mean squared errors (RMSE) based on the N samples, and defined as

$$\text{ABias}(\eta, k_n) := \left| \frac{1}{N} \sum_{i=1}^N \frac{\hat{\eta}^{(i)}}{\eta} - 1 \right| \quad \text{and} \quad \text{RMSE}(\eta, k) := \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{\hat{\eta}^{(i)}}{\eta} - 1 \right)^2},$$

where η is either γ or $x(p)$, and $\hat{\eta}^{(i)}$ is the i -th value ($i = 1, \dots, N$) of an estimator of γ or $x(p)$ evaluated at k_n (see [4], Section 6.2).

In Figure 1, resp. Figure 2, we plot the results against the corresponding k_n values of the tail index estimators, resp. the extreme quantile estimators, for each of the models by row, and by column, the ABias (*left*) and RMSE (*right*). We have four curves in each graph. The full line corresponds to $\hat{\gamma}_{k_n}^{(H)}$ (resp. $\hat{x}_{k_n}^{(W)}(p)$) estimator, the dashed line to $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$ (resp. $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}(p)$) estimator, the dotted line to $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}$ (resp. $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}(p)$) estimator and the longdashed line to $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$ (resp. $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$) estimator.

By these simulations, we observe that:

1. The Figure 1 shows that our tail index estimators ($\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}$ and $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$) look the same as the $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$ estimator given in [4], in term of ABias and RMSE for the different Models. In Models 1, 2 and 3 especially, we observe a longer stability of the three asymptotically unbiased estimators as a function of k_n , which is not the case for the Hill's estimator $\hat{\gamma}_{k_n}^{(H)}$. Our first estimator $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}$ sticks at a lower level for that. However, the Hill's estimator shows better performance for small values of k_n . Now, regarding the *GARCH*(1, 1) model, the Hill's estimator performs very poorly, whereas our two unbiased estimators $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}$ and $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$ are at least as good as $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$ estimator with the best performance for $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$ estimator.
2. In Figure 2, our extreme quantile estimators ($\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}(p)$ and $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$) follow the same profile as $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}(p)$ (introduced in [4]), in term of ABias and RMSE for the different Models with a longer stability as a function of k_n , better than the Weissman's estimator ($\hat{x}_{k_n}^{(W)}(p)$). In Models 1, 2 and 3, we observe that our unbiased estimator $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}(p)$ remains at a lower level. However, the Weissman's estimator shows better performance for small values of k_n . In Model 4, our unbiased estimator $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$ turns out best compared to the others.

In summary, this simulation study points out that the Generalized Jackknife methodology which consists of removing the bias, the extreme value index and extreme quantile estimators remain stable for a broad range of k_n values even if the dataset exhibits serial dependence. Therefore, the bias reduction procedure related to the serial dependence case enables to overcome the problem of the choice of the intermediate sequence k_n used in the application of extreme value statistics to financial time series.

5. APPLICATION TO FINANCIAL MARKET INDEX

In this section, we illustrate the performance of our bias reduction procedure using a real-world case study of the S&P500 index financial data. We estimate the tail index and extreme quantiles estimators to evaluate the S&P500 index ($I_t, t = 1, \dots, n$). Figure 3 shows the daily negative log-returns $R_t = \log(I_t/I_{t-1})$ for negative values (*Loss*), for $n = 1000$ values of S&P500 index from *April 23th*, 2018 to *April 8th*, 2022. In terms of risk management, Value-at-Risk (VaR) is a commonly used quantity in the international regulatory framework, known as the Basel Accords (see [25], for a historical review of the Basel International Settlement). The Basel Accord requires the largest international banks to hold regulatory capital for the trading book based on a 99%-VaR over a 1-day or 10-day holding period. The VaR-based risk capital calculation has attracted significant attention over the last two decades (see [21], Chapter 1). The α -VaR for the horizon $h = 1$ day is the quantile $x(p) = U(1/p)$ such that $p = 1 - \alpha$ of the distribution for the index daily log-returns.

From Figure 3, we observe that although the loss return series can be regarded as stationary, there is evidence of serial dependence such as cluster of volatility. To model this kind of phenomenon, the following *GARCH*(1, 1) model (5.17) is appropriate :

$$\begin{cases} X_t = \sigma_t \varepsilon_t, \\ \sigma_t^2 = \hat{\lambda}_0 + \hat{\lambda}_1 X_{t-1}^2 + \hat{\lambda}_2 \sigma_{t-1}^2, \end{cases} \tag{5.17}$$

Furthermore, we fit the *GARCH*(1,1) model to our dataset, where the innovations ε_t are independent and identically Student- t distributed with $\hat{\lambda}_0 = 2.3 \times 10^{-05}$ (2.657×10^{-05}), $\hat{\lambda}_1 = 7.574 \times 10^{-01}$ (8.665×10^{-01}), $\hat{\lambda}_2 = 7.037 \times 10^{-01}$ (6.539×10^{-02}) and the parameter of the Student- t is $\hat{\nu} = 2.189$ (2.453×10^{-01}), where the value in parentheses is the standard deviation. Our aim is to estimate the Value-at-Risk of the return series at the 99.9% level, which corresponds to a extreme quantile with tail probability 0.1%, *i.e.*, $x(0.001)$.

Following the estimation procedure, we firstly estimate the second order parameter $\hat{\rho}_{k_\rho} = -1.468$ which corresponds to the value of $k_\rho = 442$. Secondly, we use the four estimators which are $\hat{\gamma}_{k_n}^{(H)}$, $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}$, $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}$ and $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$ to estimate the tail index of the loss return series. We see that our $\hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}$ estimator seems to be more stable while the others increase with k_n . Finally, we use $\hat{x}_{k_n}^{(W)}(p)$, $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}(p)$, $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}(p)$, and $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$ to estimate the VaR at 99% and 99.9% levels. It is clear that the $\hat{x}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub2)}(p)$ quantile estimator is more stable than the others regarding the middle and right hand side Figure 4. That means it is the best estimator which can be used for high quantile estimation.

To illustrate that, we proceed by evaluating the optimal value of k_n that we denote by k_n^* . In the figure, this is the point around which a tail index estimator $\hat{\gamma}_{k_n}^{(\bullet)}$ is stable. It is essential to decide a value of k_n which will be used to get the high quantile estimator. The selection of k_n^* is equivalent to the choice of the threshold in the EVT peaks-over-threshold method. An alternative method is the algorithm of [23], p. 137, which allows to get an automatic choice of the number of top extremes k_n for a given tail index estimator $\hat{\gamma}_{k_n}^{(\bullet)}$. According to these authors, an automatic choice of k_n^* is the k_n value which minimizes

$$\frac{1}{k_n} \sum_{j=1}^{k_n} j^\tau \left| \hat{\gamma}_j^{(\bullet)} - \text{med} \left(\hat{\gamma}_1^{(\bullet)}, \dots, \hat{\gamma}_{k_n}^{(\bullet)} \right) \right|, \quad 1 \leq k_n \leq N,$$

where $0 \leq \tau < 1/2$ and $\text{med} \left(\hat{\gamma}_1^{(\bullet)}, \dots, \hat{\gamma}_{k_n}^{(\bullet)} \right)$ denotes the median of $\left(\hat{\gamma}_1^{(\bullet)}, \dots, \hat{\gamma}_{k_n}^{(\bullet)} \right)$ and N is the size of the negative log-return series.

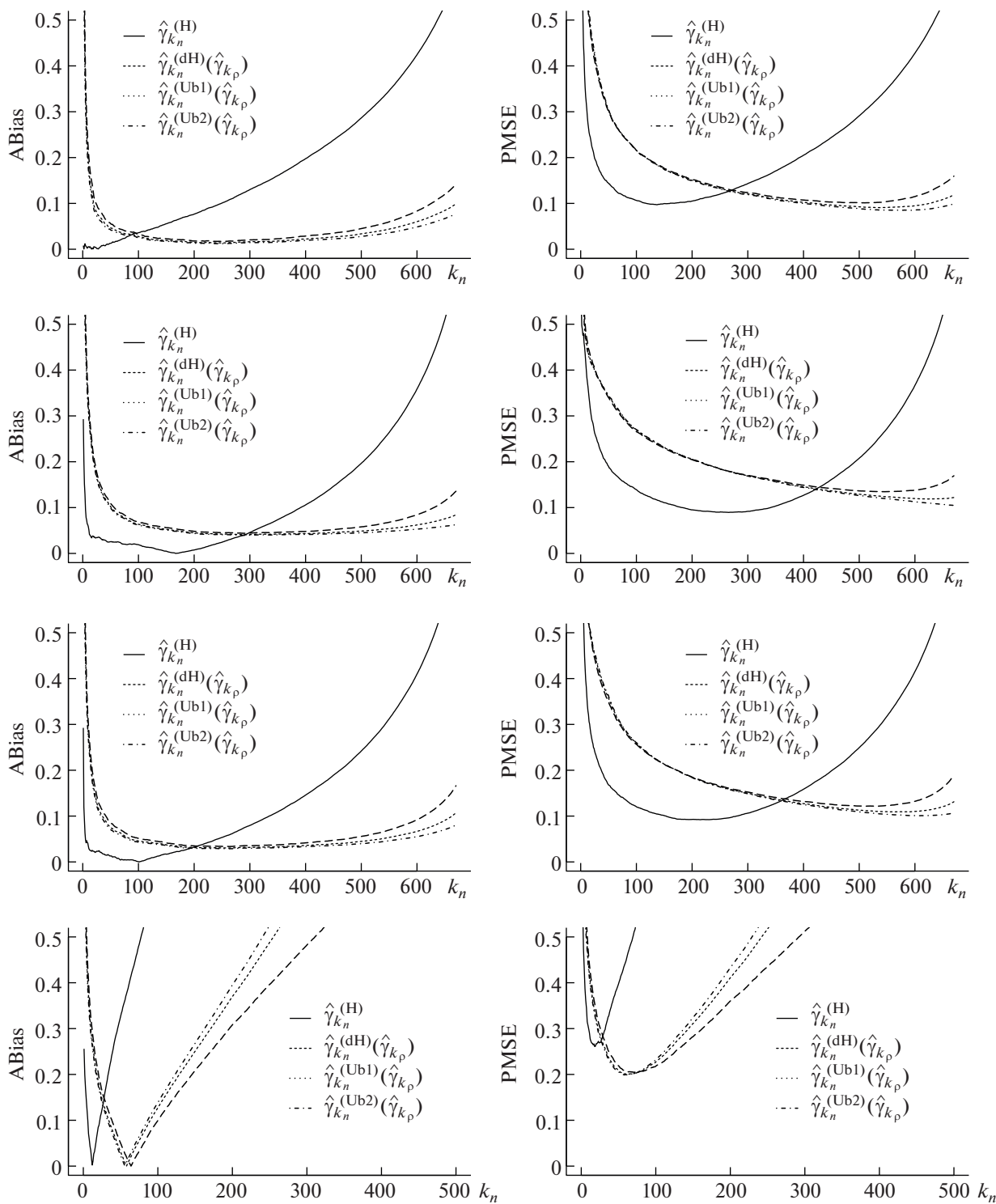


Fig. 1. Simulation of the tail index: By row, **Models 1, 2, 3, 4.** By column, ABias (*left*) and RMSE (*right*) as functions of k_n .

By the way, choosing $\tau = 1/4$, we compute the optimal values k_n^* associated to the estimated values $\hat{\gamma}_{k_n}^{(\bullet)} \in \left\{ \hat{\gamma}_{k_n}^{(H)}, \hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)}, \hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(dH)}, \hat{\gamma}_{k_n, \hat{\rho}_{k_\rho}}^{(Ub1)} \right\}$. In Table 1 below, we present the results of the estimated values $\hat{\gamma}_{k_n^*}^{(\bullet)}$ with their associated extreme quantile estimators. Since we do not employ a parametric model for

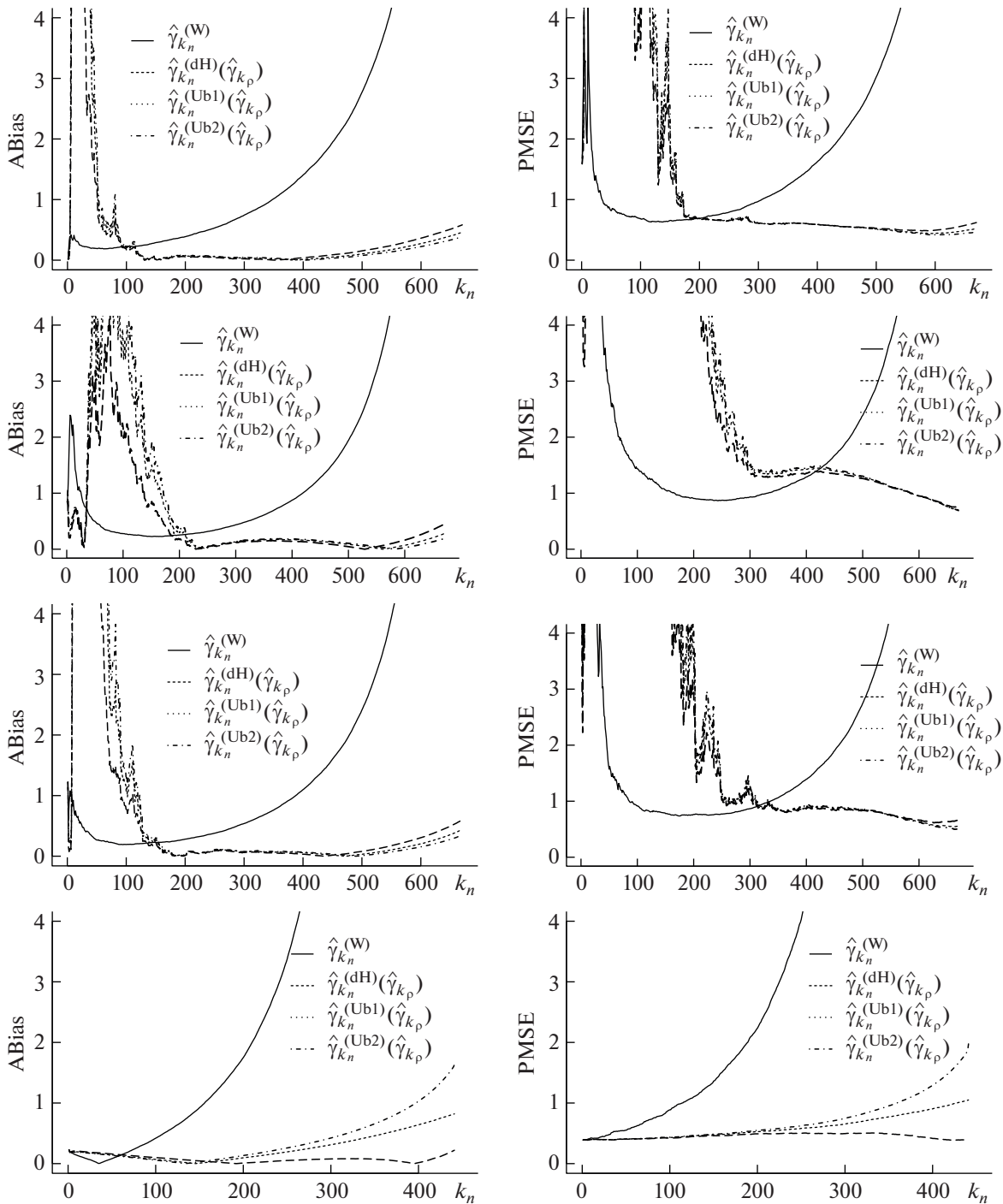


Fig. 2. Simulation of the extreme quantiles: By row, **Models 1, 2, 3, 4.** By column, ABias (*left*) and RMSE (*right*) as functions of k_n .

the time series, there is no explicit formula for calculating the asymptotic variance of the two estimators. Therefore, we opt to use a block bootstrapping method to construct the confidence interval for extreme quantiles estimators. The block bootstrapping follows the routine `boot` in the package `boot` in R software. By repeating such a bootstrapping procedure $K = 1000$ times, we obtain K bootstrapped estimates for each estimator. The sample standard deviation across the K estimates gives an estimate of the standard deviation of the underlying estimator for given $k_n \in \{1, \dots, N\}$. We construct the 95% confidence interval using the point estimate and the estimated standard deviation. This procedure is

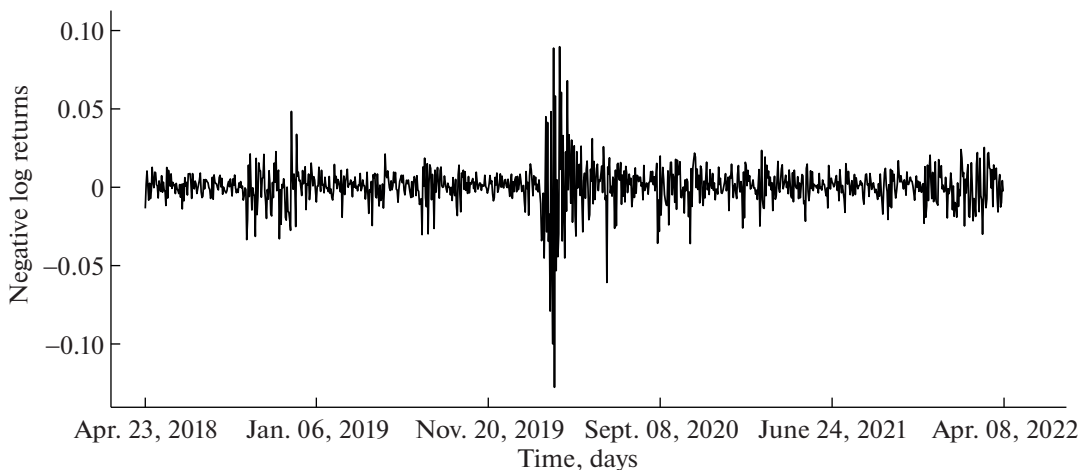


Fig. 3. S&P500 index data: daily negative log-returns from April 23th, 2018 to April 8th, 2022.

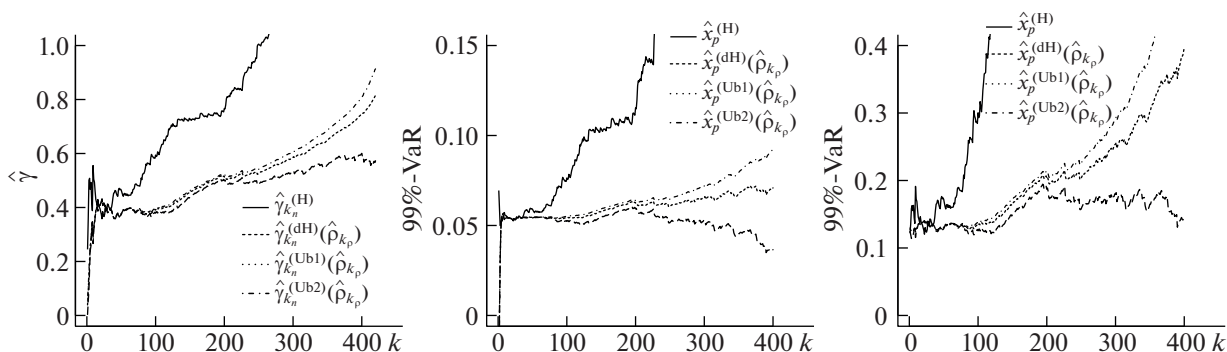


Fig. 4. S&P500 index data: estimated values of γ (left panel), 99%-VaR (middle panel) and 99.9%-VaR (right panel) as function of k_n using the four estimators.

applied to all values of k_n of each estimator. The point estimates of the 99% and 99.9% Value at Risk as well as the lower and upper bounds of the confidence intervals are given in Table 1.

From Table 1, we remark that for moderate quantiles ($p = 0.01$), our $\hat{x}_{k_n^*, \hat{\rho}_{k_p}}^{(Ub2)}(p)$ estimator is not very competitive regarding the cover values. In contrast, for high quantiles ($p = 0.001$), our estimator is the best compared to that $\hat{x}_{k_n^*, \hat{\rho}_{k_p}}^{(dH)}(p)$. That illustrates well our conclusions drawn from the graphical analysis.

6. CONCLUSION

In this paper, we have presented three improved estimators of the tail index and extreme quantiles for stationary β -mixing data. The proposed estimators are based on the generalized Jackknife methodology and rely on taking any pair of three special cases of our family of estimators to cancel the bias term. We have shown that our proposed estimators show good performance compared to existing methods in terms of accuracy and stability over k . We have also conducted a thorough simulation study to demonstrate the effectiveness of our proposed method in practice. Our results have important implications for risk management and decision-making in various industries where accurate estimation of extreme events is crucial. Overall, our work contributes to the ongoing efforts in developing reliable and efficient methods for extreme value analysis of dependent data.

7. PROOFS OF THE MAIN RESULTS

Proof of Proposition 2.1 It is similar to that of Colorally A.2 in [4] where we limit ourselves only to the second order conditions.

Assume that the assumptions in Proposition 2.1 hold. With respect to the same Skorohod construction as in Proposition 1 in [3], it results that, for given $\alpha, \delta > 0$, there exist a function $\tilde{A} \sim A$, and a centered Gaussian process $(W(t))_{t \in [0,1]}$ with covariance function r , such that, as $n \rightarrow \infty$:

$$\sup_{t \in (0,1)} t^{1/2+\delta} \left| \sqrt{k_n} \left(\log \frac{Q_n(t)}{Q_n(1)} - \gamma(-\log t) \right) - \gamma(t^{-1}W(t) - W(1)) - \sqrt{k_n} \tilde{A}(n/k_n) \frac{t^{-\rho} - 1}{\rho} \right| \rightarrow 0 \text{ a.s.}$$

This means for all $t \in (0, 1]$

$$\left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha = \gamma^\alpha(-\log t)^\alpha \left[1 + \frac{(-\log t)^{-1}}{\sqrt{k_n}} (t^{-1}W(t) - W(1)) + \gamma^{-1}(-\log t)^{-1} \tilde{A}(n/k_n) \frac{t^{-\rho} - 1}{\rho} + o(k_n^{-1/2})t^{-1/2-\delta} \right]^\alpha.$$

And from the relation $(1 + x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2}x^2 + o(x^2), o(x^2) \rightarrow 0$ when $x \rightarrow 0$, we obtain

$$\sup_{t \in (0,1)} t^{1/2+\delta} \left| \sqrt{k_n} \left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha - \gamma^\alpha(-\log t)^\alpha - \alpha \gamma^\alpha(-\log t)^{\alpha-1} (t^{-1}W(t) - W(1)) - \sqrt{k_n} \tilde{A}(n/k_n) \alpha \gamma^{\alpha-1}(-\log t)^{\alpha-1} \frac{t^{-\rho} - 1}{\rho} \right| \rightarrow 0 \text{ a.s.}$$

Without losing generality, some terms tend to 0, as $n \rightarrow \infty$. In fact, we have $\sup_{t \in (0,1)} t^{1/2+\epsilon} t^{-1} |W(t)| = O(1)$ a.s. and $\tilde{A}(n/k_n) \rightarrow 0$ as $n \rightarrow \infty$.

Proof of Corollary 2.1 Obviously, we have

$$M_k^{(\alpha)} = \int_0^1 \left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha dt \tag{7.19}$$

By using $\delta < 1/2$ from the Proposition 2.1, we can take the integral of $\left(\log \frac{Q_n(t)}{Q_n(1)} \right)^\alpha$ on $(0, 1]$ and use the fact that $\int_0^1 (-\log t)^{a-1} t^{-b} dt = \frac{\Gamma(a)}{(1-b)^a}$ for $b < 1$ to obtain the result in the Corollary. The random term is obtained by taking

$$P^{(\alpha)} = \int_0^1 (-\log t)^{\alpha-1} (t^{-1}W(t) - W(1)) dt.$$

and is normally distributed with mean zero and covariance $Cov(P^{(\alpha)}, P^{(\beta)}) = c_{\alpha,\beta}$ defined as,

$$c_{\alpha,\beta} = \int_0^1 \int_0^1 (-\log s)^{\alpha-1} (-\log t)^{\beta-1} \times \left(\frac{r(s,t)}{st} - \frac{r(s,1)}{s} - \frac{r(1,t)}{t} + r(1,1) \right) ds dt,$$

with the covariance structure r defined as in the regularity condition (C_R) .

Proof of Theorem 2.1. In accordance with the assumptions outlined in Corollary 2.1, get the following representation:

$$M_{k_n}^{(\alpha)} \stackrel{d}{=} \gamma^\alpha \Gamma(\alpha + 1) + \frac{\alpha \gamma^\alpha P^{(\alpha)}}{\sqrt{k_n}} + \tilde{A}(n/k_n) \gamma^{\alpha-1} \frac{\Gamma(\alpha + 1)}{\rho} \left(\frac{1}{(1-\rho)^\alpha} - 1 \right) + o_{\mathbb{P}}(k_n^{-1/2}) \tag{7.20}$$

It follows

$$\frac{M_{k_n}^{(\alpha)}}{\Gamma(\alpha + 1)} \stackrel{d}{=} \gamma^\alpha \left(1 + \frac{\alpha P^{(\alpha)}}{\sqrt{k_n} \Gamma(\alpha + 1)} + \tilde{A}(n/k_n) \frac{\gamma^{-1}}{\rho} \left(\frac{1}{(1 - \rho)^\alpha} - 1 \right) + o_{\mathbb{P}}(k_n^{-/2}) \right). \tag{7.21}$$

Therefore

$$\tilde{\gamma}_{k_n}^{(\alpha)} := \left(\frac{M_{k_n}^{(\alpha)}}{\Gamma(\alpha + 1)} \right)^{\frac{1}{\alpha}} \stackrel{d}{=} \gamma \left(1 + \frac{\alpha P^{(\alpha)}}{\sqrt{k_n} \Gamma(\alpha + 1)} + \tilde{A}(n/k_n) \frac{\gamma^{-1}}{\rho} \left(\frac{1}{(1 - \rho)^\alpha} - 1 \right) + o_{\mathbb{P}}(k_n^{-/2}) \right)^{\frac{1}{\alpha}}.$$

Using a Taylor expansion, we get as $n \rightarrow \infty$,

$$\begin{aligned} \tilde{\gamma}_{k_n}^{(\alpha)} &\stackrel{d}{=} \gamma \left(1 + \frac{1}{\alpha} \left[\frac{\alpha P^{(\alpha)}}{\sqrt{k_n} \Gamma(\alpha + 1)} + \tilde{A}(n/k_n) \frac{\gamma^{-1}}{\rho} \left(\frac{1}{(1 - \rho)^\alpha} - 1 \right) \right] + o_{\mathbb{P}}(k_n^{-/2}) \right) \\ &\stackrel{d}{=} \gamma + \frac{\gamma P^{(\alpha)}}{\sqrt{k_n} \Gamma(\alpha + 1)} + \frac{\tilde{A}(n/k_n)}{\alpha \rho} \left(\frac{1}{(1 - \rho)^\alpha} - 1 \right) + o_{\mathbb{P}}(k_n^{-/2}). \end{aligned}$$

The expression (2.5) of $\hat{\gamma}_{k_n}^{(\alpha)}$ can be written as:

$$\hat{\gamma}_{k_n}^{(\alpha)} = \frac{M_{k_n}^{(\alpha)}}{\Gamma(\alpha + 1)} \left(M_{k_n}^{(1)} \right)^{1-\alpha}, \quad \alpha > 1.$$

From (7.20), it follows:

$$\begin{aligned} M_{k_n}^{(1)} &\stackrel{d}{=} \gamma + \frac{\gamma P^{(1)}}{\sqrt{k_n}} + \frac{\tilde{A}(n/k_n)}{1 - \rho} + o_{\mathbb{P}}(k_n^{-/2}) \\ &\stackrel{d}{=} \gamma \left[1 + \frac{P^{(1)}}{\sqrt{k_n}} + \tilde{A}(n/k_n) \frac{\gamma^{-1}}{1 - \rho} + o_{\mathbb{P}}(k_n^{-/2}) \right]. \end{aligned}$$

$$\begin{aligned} \left(M_{k_n}^{(1)} \right)^{1-\alpha} &\stackrel{d}{=} \gamma^{1-\alpha} \left(1 + \frac{P^{(1)}}{\sqrt{k_n}} + \tilde{A}(n/k_n) \frac{\gamma^{-1}}{1 - \rho} + o_{\mathbb{P}}(k_n^{-/2}) \right)^{1-\alpha} \\ &\stackrel{d}{=} \gamma^{1-\alpha} \left(1 + \frac{(1 - \alpha) P^{(1)}}{\sqrt{k_n}} + \tilde{A}(n/k_n) \gamma^{-1} \frac{1 - \alpha}{1 - \rho} + o_{\mathbb{P}}(k_n^{-/2}) \right). \end{aligned}$$

We recall that

$$\frac{M_{k_n}^{(\alpha)}}{\Gamma(\alpha + 1)} \stackrel{d}{=} \gamma^\alpha \left(1 + \frac{\alpha P^{(\alpha)}}{\sqrt{k_n} \Gamma(\alpha + 1)} + \tilde{A}(n/k_n) \frac{\gamma^{-1}}{\rho} \left(\frac{1}{(1 - \rho)^\alpha} - 1 \right) + o_{\mathbb{P}}(k_n^{-/2}) \right).$$

Hence

$$\hat{\gamma}_{k_n}^{(\alpha)} \stackrel{d}{=} \gamma + \frac{\gamma}{\sqrt{k_n}} \hat{P}^{(\alpha)} + \tilde{A}(n/k_n) \left\{ \frac{1 - (1 - \rho)^\alpha}{\rho(1 - \rho)^\alpha} - \frac{\alpha - 1}{1 - \rho} \right\} + o_{\mathbb{P}}(k_n^{1/2}).$$

Where $\hat{P}^{(\alpha)} = \frac{\alpha P^{(\alpha)}}{\Gamma(\alpha + 1)} - (\alpha - 1) P^{(1)}$, which ends the proof of the Theorem 2.1.

Proof of Corollary 2.2. The results in Corollary 2.2 quite direct by computing the variance of the centered Gaussian term $P^{(\alpha)}$ given in Corollary 2.1.

Proof of Theorem 2.2. Note That:

$$\hat{\gamma}_{k_n}^{(H)} = \tilde{\gamma}_{k_n}^{(1)} = \hat{\gamma}_{k_n}^{(1)} = M_{k_n}^{(1)}, \quad \hat{\gamma}_{k_n}^{(2)} = \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \quad \text{and} \quad \tilde{\gamma}_{k_n}^{(2)} = \sqrt{\frac{M_{k_n}^{(2)}}{2}}. \tag{7.22}$$

We recall that:

$$\widehat{\gamma}_{k_n, \widehat{\rho}}^{(Ub1)} := \frac{1}{\widehat{\rho}} \left(\sqrt{2M_{k_n}^{(2)}} - (2 - \widehat{\rho}) \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \right),$$

$$\widehat{\gamma}_{k_n, \widehat{\rho}}^{(Ub2)} := \frac{1}{\widehat{\rho}} \left((2 - \widehat{\rho})M_{k_n}^{(1)} - (1 - \widehat{\rho})\sqrt{2M_{k_n}^{(2)}} \right)$$

and

$$\widehat{\gamma}_{k_n, \widehat{\rho}}^{(dH)} := \frac{1}{\widehat{\rho}} \left(M_{k_n}^{(1)} - (1 - \widehat{\rho}) \frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \right),$$

where $\widehat{\rho}$ is either a canonical negative value $\widehat{\rho} := \rho = \rho_0$ or a consistent estimator $\widehat{\rho} := \widehat{\rho}_{k_\rho}$ of ρ , with $k_\rho := k_{\rho, n}$ an intermediate sequence of integers greater than k_n , satisfying $k_\rho \rightarrow \infty$ and $k_\rho/n \rightarrow 0$, as $n \rightarrow \infty$.

From Theorem 2.1, we have:

$$M_{k_n}^{(1)} \stackrel{d}{=} \gamma + \frac{\gamma P^{(1)}}{k_n^{1/2}} + \frac{\widetilde{A}(n/k_n)}{1 - \rho} + o_{\mathbb{P}}(k_n^{-1/2}), \tag{7.23}$$

$$\sqrt{2M_{k_n}^{(2)}} \stackrel{d}{=} 2\sqrt{\frac{M_{k_n}^{(2)}}{2}} \stackrel{d}{=} 2\gamma + \frac{\gamma P^{(2)}}{k_n^{1/2}} + \widetilde{A}(n/k_n) \frac{2 - \rho}{(1 - \rho)^2} + o_{\mathbb{P}}(k_n^{-1/2}), \tag{7.24}$$

and

$$\frac{M_{k_n}^{(2)}}{2M_{k_n}^{(1)}} \stackrel{d}{=} \gamma + \frac{\gamma P^{(2)}}{k_n^{1/2}} - \frac{\gamma P^{(1)}}{k_n^{1/2}} + \widetilde{A}(n/k_n) \frac{1}{(1 - \rho)^2} + o_{\mathbb{P}}(k_n^{-1/2}). \tag{7.25}$$

Using the Jackknife methodology, this leads to:

$$\widehat{\gamma}_{k_n, \widehat{\rho}}^{(\bullet)} \stackrel{d}{=} \gamma + \frac{\gamma(2 - \widehat{\rho})P^{(1)} - \gamma(1 - \widehat{\rho})P^{(2)}}{\widehat{\rho}\sqrt{k_n}} + \widetilde{A}(n/k) \frac{\widehat{\rho} - \rho}{\widehat{\rho}(1 - \rho)^2} + o_{\mathbb{P}}(k_n^{-1/2}).$$

For a given canonical negative value $\widehat{\rho} := \rho = \rho_0$ or a consistent estimator $\widehat{\rho} := \widehat{\rho}_{k_\rho}$ of ρ , we get as $n \rightarrow \infty$

$$k_n^{1/2} \left(\widehat{\gamma}_{k_n, \widehat{\rho}}^{(\bullet)} - \gamma \right) \xrightarrow{d} \frac{\gamma}{\rho} \left((2 - \rho)P^{(1)} - (1 - \rho)P^{(2)} \right).$$

Computing the the variance term of $\frac{\gamma}{\rho} \left((2 - \rho)P^{(1)} - (1 - \rho)P^{(2)} \right)$. with respect to the covariance structure, the Theorem 2.2 holds.

Proof of Corollary 2.2. According to [4], under (C_{SO}) and (C_R) assumptions, if $k_\rho := k_{\rho, n}$ satisfies $k_\rho \rightarrow \infty$, $k_\rho/n \rightarrow 0$ and $\sqrt{k_\rho}A(n/k_\rho) \rightarrow \infty$, as $n \rightarrow \infty$, then $\widehat{\rho}_{k_\rho}^{(*)} \xrightarrow{\mathbb{P}} \rho$. The Corollary 2.3 follows by applying the Theorem 2.2.

Proof of Theorem 2.3. For simplify, let's denote by $d_n = \frac{k_n}{np}$ and

$$T_n = \frac{\left(M_{k_n}^{(2)} - 2(M_{k_n}^{(1)})^2 \right) (1 - \widehat{\rho})^2}{2M_{k_n}^{(1)} \widehat{\rho}^2} \left(1 - d_n^{\widehat{\rho}} \right).$$

We consider the expression:

$$E_n = \frac{\sqrt{k_n}}{\log d_n} \left(\frac{\widehat{x}_{k_n, \widehat{\rho}}^{(\bullet)}(p)}{x(p)} - 1 \right) \quad \text{with} \quad \widehat{x}_{k_n, \widehat{\rho}}^{(\bullet)}(p) = X_{n-k_n, n} d_n^{\widehat{\rho}} (1 - T_n)$$

. Let $x(p) = U(1/p)$. It follows

$$\begin{aligned}
E_n &= \frac{\sqrt{k_n}}{\log d_n} \left(\frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)}} (1 - T_n)}{x_p} - 1 \right) \\
&= \frac{\sqrt{k_n}}{\log d_n} \left[\left(\frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)}}}{x(p)} - 1 + 1 \right) (1 - T_n) - 1 \right] \\
&= \frac{\sqrt{k_n}}{\log d_n} \left[\left(\frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)}}}{x(p)} - 1 \right) (1 - T_n) + (1 - T_n) - 1 \right] \\
&= \frac{\sqrt{k_n}}{\log d_n} \left(\frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)}}}{x(p)} - 1 \right) (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} T_n \\
&= \frac{\sqrt{k_n}}{\log d_n} \frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)}}}{x(p)} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} T_n \\
&= \frac{\sqrt{k_n}}{\log d_n} \frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma + \gamma} U(n/k_n)}{x(p) U(n/k_n)} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} T_n \\
&= \frac{d_n^\gamma U(n/k_n)}{U(1/p_n)} \frac{\sqrt{k_n}}{\log d_n} \frac{X_{n-k_n, n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma}}{U(n/k_n)} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} T_n \\
&= \frac{d_n^\gamma U(n/k_n)}{U(1/p_n)} \left[\frac{\sqrt{k_n}}{\log d_n} \left(\frac{X_{n-k_n, n}}{U(n/k_n)} - 1 \right) d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma} + \frac{\sqrt{k_n}}{\log d_n} d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma} \right] (1 - T_n) \\
&\quad - \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} T_n \\
&= \frac{d_n^\gamma U(n/k_n)}{U(1/p_n)} \left[\frac{\sqrt{k_n}}{\log d_n} \left(\frac{X_{n-k_n, n}}{U(n/k_n)} - 1 \right) d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma} + \frac{\sqrt{k_n}}{\log d_n} \left(d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma} - 1 \right) \right] (1 - T_n) \\
&\quad + \frac{d_n^\gamma U(n/k_n)}{U(1/p_n)} \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n}.
\end{aligned}$$

We have as $n \rightarrow \infty$, $\frac{d_n^\gamma U(n/k_n)}{U(1/p_n)} \rightarrow 1$, because $\frac{U(1/p_n) d_n^{-\gamma}}{U(n/k_n)} \rightarrow 1$, by the regularly varying condition of the tail quantile U .

From Proposition 1. in [3], we have as $n \rightarrow \infty$,

$$\sqrt{k} \left(\frac{X_{n-k, n}}{U(n/k)} - 1 \right) \xrightarrow{d} \gamma W(1). \quad (7.26)$$

Next, we also

$$d_n^{\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma} = \exp \left[\left(\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma \right) \log d_n \right].$$

From Theorem 2.2, we get for all large values of n ;

$$\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma \stackrel{d}{=} k_n^{-1/2} \Gamma + o_{\mathbb{P}}(k_n^{-1/2})$$

as $n \rightarrow \infty$ where Γ is a centred Gaussian distribution with variance by assumption $\sigma^2(\gamma, \rho)$. Since $\log d_n/k_n^{1/2} \rightarrow 0$, as $n \rightarrow \infty$, this implies that for all n large enough,

$$\left(\widehat{\gamma}_{k_n, \rho}^{(\bullet)} - \gamma \right) \log d_n = o_{\mathbb{P}}(1). \quad (7.27)$$

Further, by using the inequality, $\left| \frac{\exp(x)-1}{x} - 1 \right| \leq \exp(|x|) - 1$, in the neighborhood of zero, we get

$$d_n^{\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet) - \gamma}} \stackrel{d}{=} 1 + \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet)} - \gamma \right) \log d_n + O_{\mathbb{P}}(R_n), \tag{7.28}$$

as $n \rightarrow \infty$, with

$$R_n = \left| \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet)} - \gamma \right) \log d_n \right| \times \left[\exp \left\{ \left| \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet)} - \gamma \right) \log d_n \right| \right\} - 1 \right].$$

Since $np_n/k_n \rightarrow 0$, as $n \rightarrow \infty$, we have $d_n = k_n/(np_n) \rightarrow \infty$. Therefore, from (7.26), (7.27) and (7.28), we get as $n \rightarrow \infty$:

$$\frac{\sqrt{k_n}}{\log d_n} \left(\frac{X_{n-k_n, n}}{U(n/k_n)} - 1 \right) d_n^{\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet) - \gamma}} = o_{\mathbb{P}}(1).$$

And in the other hand, we get as $n \rightarrow \infty$:

$$\frac{\sqrt{k_n}}{\log d_n} \left(d_n^{\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet) - \gamma}} - 1 \right) \stackrel{d}{=} \sqrt{k_n} \left(\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet)} - \gamma \right) + o_{\mathbb{P}}(1).$$

This leads to

$$\frac{\sqrt{k_n}}{\log d_n} \left(d_n^{\hat{\gamma}_{k_n, \hat{\rho}}^{(\bullet) - \gamma}} - 1 \right) \xrightarrow{d} \Gamma,$$

as $n \rightarrow \infty$. Next, from (7.20), $M_{k_n}^{(1)} \xrightarrow{\mathbb{P}} \gamma$ and $M_{k_n}^{(2)} \xrightarrow{\mathbb{P}} 2\gamma^2$, as $n \rightarrow \infty$. For a given canonical negative value $\hat{\rho} := \rho = \rho_0$ or a consistent estimator $\hat{\rho} := \hat{\rho}_{k_\rho}$ of ρ , we get $T_n \xrightarrow{\mathbb{P}} 0$, as $n \rightarrow \infty$.

For the last term, we have:

$$\begin{aligned} F_n &= \frac{d_n' U(n/k_n)}{U(1/p_n)} \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} \\ &= \frac{\sqrt{k_n}}{\log d_n} \frac{d_n' U(n/k_n)}{U(1/p_n)} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} \\ &= \frac{\sqrt{k_n}}{\log d_n} \left[\frac{d_n' U(n/k_n)}{U(1/p_n)} - 1 \right] (1 - T_n) + \frac{\sqrt{k_n}}{\log d_n} (1 - T_n) - \frac{\sqrt{k_n}}{\log d_n} \\ &= \frac{\sqrt{k_n}}{\log d_n} \left[\frac{d_n' U(n/k_n)}{U(1/p_n)} - 1 \right] (1 - T_n) + \frac{\sqrt{k_n}}{\log d_n} T_n \\ &= F_{n,1} + F_{n,2}. \end{aligned}$$

By expanding $F_{n,1}$ we get:

$$\begin{aligned} F_{n,1} &= \frac{\sqrt{k}}{\log d_n} \frac{d_n' U(n/k_n)}{U(1/p_n)} \left[1 - \frac{d_n^{-\gamma} U(1/p_n)}{U(n/k)} \right] (1 - T_n) \\ &= \frac{1}{\log d_n} \frac{-d_n' U(n/k)}{U(1/p_n)} \frac{\sqrt{k} A(n/k)}{A(n/k)} \left[\frac{d_n^{-\gamma} U(1/p_n)}{U(n/k)} - 1 \right] (1 - T_n). \end{aligned}$$

Since $\frac{-d_n' U(n/k)}{U(1/p_n)} \rightarrow 1$, $\sqrt{k} A(n/k) \rightarrow \lambda$, $\log d_n \rightarrow \infty$, $\frac{d_n^{-\gamma} U(1/p_n)}{A(n/k)} - 1 \rightarrow -\frac{1}{\rho}$, and $T_n \xrightarrow{\mathbb{P}} 0$, we get $F_{n,1} \xrightarrow{\mathbb{P}} 0$, as $n \rightarrow \infty$.

Now, for $F_{n,2}$, we obtain:

$$F_{n,2} = \frac{\sqrt{k_n}}{\log d_n} \left(T_n - \frac{\tilde{A}(n/k_n)}{\rho} (1 - d_n^{\hat{\rho}}) \right) + \frac{\sqrt{k_n}}{\log d_n} \frac{\tilde{A}(n/k_n)}{\rho} (1 - d_n^{\hat{\rho}}).$$

For a given canonical negative value $\hat{\rho} := \rho = \rho_0$ or a consistent estimator $\hat{\rho} := \hat{\rho}_{k_p}$ of ρ , we get $d_n^{\hat{\rho}} \rightarrow 0$, as $n \rightarrow \infty$. Since $\sqrt{k_n}A(n/k_n) \rightarrow \lambda$, $\log d_n \rightarrow \infty$, we have, as $n \rightarrow \infty$:

$$\frac{\sqrt{k_n} \tilde{A}(n/k_n)}{\log d_n} \frac{1}{\rho} (1 - d_n^{\hat{\rho}}) \rightarrow 0.$$

On the other hand

$$\frac{\sqrt{k_n}}{\log d_n} \left(T_n - \frac{\tilde{A}(n/k_n)}{\rho} (1 - d_n^{\hat{\rho}}) \right) = \frac{\sqrt{k_n}}{\log d_n} \left(\frac{(M_{k_n}^{(2)} - 2(M_{k_n}^{(1)})^2) (1 - \hat{\rho})^2}{2M_{k_n}^{(1)} \hat{\rho}^2} - \frac{\tilde{A}(n/k_n)}{\rho} \right) (1 - d_n^{\hat{\rho}}).$$

From (7.23) and (7.25), we have:

$$\sqrt{k_n} \left(\frac{M_{k_n}^{(2)} - 2(M_{k_n}^{(1)})^2}{2\hat{\rho}M_{k_n}^{(1)}} - \frac{\tilde{A}(n/k_n)}{\rho} \right) \xrightarrow{d} \frac{(1 - \rho)^2}{\rho^2} (P^{(2)} - 2P^{(1)}).$$

This implies that

$$\frac{\sqrt{k_n}}{\log d_n} \left(T_n - \frac{\tilde{A}(n/k_n)}{\rho} \right) \xrightarrow{\mathbb{P}} 0.$$

Finally, we get, as $n \rightarrow \infty$,

$$\frac{\sqrt{k_n}}{\log d_n} \left(\frac{\hat{x}_{k_n, \hat{\rho}}^{(\bullet)}(p)}{x(p)} - 1 \right) \xrightarrow{d} \Gamma.$$

This ends the proof of Theorem 2.3.

ACKNOWLEDGEMENTS

The authors would like to thank the Reviewers and editors for their valuable comments and suggestions that helped improve considerably the paper.

REFERENCES

1. M. I. Fraga Alves, M. Ivette Gomes, and Laurens de Haan, "A new class of semiparametric estimators of the second order parameter," In: *Portugaliae Mathematica* **60** (9), 193–214 (2003).
2. Jan Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys, "Tail Index Estimation and an Exponential Regression Model," In: *Extremes* **2** (2), 177–200 (1999).
3. Valérie Chavez-Demoulin and Armelle Guillou, "Extreme quantile estimation for β -mixing time series and applications," In: *Insurance: Mathematics and Economics* **83** (16), 59–74 (2018).
4. Laurens De Haan, Cécile Mercadier, and Chen Zhou, "Adapting extreme value statistics to financial time series: dealing with bias and serial dependence," In: *Finance and Stochastics* **20** (2), 321–354 (2016).
5. Arnold L. M. Dekkers, John H. J. Einmahl, and Laurens De Haan, "A moment estimator for the index of an extreme-value distribution," In: *The Annals of Statistics* **17** (3), 1833–1855.
6. El Hadji Deme, Laurent Gardes, and Stéphane Girard, "On the estimation of the second order parameter for heavy-tailed distributions," In: *REVSTAT-Statistical Journal* **11** (3), 277–299 (2013).
7. Holger Drees, "Extreme quantile estimation for dependent data, with applications to finance," In: *Bernoulli* **9** (4), 617–657 (2003).
8. Holger Drees. "Weighted approximations of tail processes for β -mixing random variables," In: *The Annals of Applied Probability* **10** (5), 1274–1301 (2000).
9. Holger Drees and Edgar Kaufmann, "Selecting the optimal sample fraction in univariate extreme value estimation," In: *Stochastic Processes and their applications* **75** (7), 149–172 (1998).
10. Andrey Feuerverger and Peter Hall, "Estimating a tail exponent by modelling departure from a Pareto distribution," In: *The Annals of Statistics* **27** (8), 760–781 (1999).
11. Yuri Goegebeur, Jan Beirlant, and Tertius de Wet, "Kernel estimators for the second order parameter in extreme value statistics," In: *Journal of statistical Planning and Inference* **140** (10), 2632–2652 (2010).

12. M Ivette Gomes and M Jo.ao Martins. "Asymptotically unbiased" estimators of the tail index based on external estimation of the second order parameter," In: *Extremes* **5** (14), 5–31 (2002).
13. M. Ivette Gomes and M. João Martins, "Generalizations of the Hill estimator.asymptotic versus finite sample behaviour," In: *Journal of statistical planning and inference* **93** (13), 161–180 (2001).
14. Bruce M. Hill, "A simple general approach to inference about the tail of a distribution," In: *The annals of statistics* **3** (17), 1163–1174 (1975).
15. Tailen Hsing, "On tail index estimation using dependent data," In: *The Annals of Statistics* **18**, 1547–1569 (1991).
16. M. Ivette Gomes, Laurens De Haan, and Lúgia Henriques Rodrigues, "Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses," In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** (1), 31–52 (2008).
17. M. Ivette Gomes, M. João Martins, and Manuela Neves, "Alternatives to a semiparametric estimator of parameters of rare events.the Jackknife methodology," In: *Extremes* **3** (12), 207–229 (2000).
18. Pavlina Jordanova, Z. Fabián, P. Hermann, et al., "Weak properties and robustness of t-Hill estimators," In: *Extremes* **19** (19), 591–626 (2016).
19. Harry Kesten, "Random difference equations and renewal theory for products of random matrices," In: **20** (1973).
20. Gunther Matthys, E. Delafosse, A. Guillou, and J. Beirlant, "Estimating catastrophic quantile levels for heavy-tailed distributions," In: *Insurance: Mathematics and Economics* **34** (21), 517–537 (2004).
21. Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts, *Quantitative risk management: concepts, techniques and tools-revised edition*, 22 (Princeton university press, 2015).
22. L. Peng, "Asymptotically unbiased estimators for the extreme-value index," In: *Statistics & Probability Letters* **38** (23), 107–115 (1998).
23. Rolf-Dieter Reiss and Michael Thomas, *Statistical Analysis of Extreme Values. With Applications to Insurance, Finance, Hydrology and Other Fields*, 24, 3rd ed. (Birkhäuser, 2007).
24. C Stărică, *On the tail empirical process of solutions of stochastic difference equations*, Tech. rep. 25 (Working paper, Chalmers University, 2000), Vol. 22.
25. Daniel K. Tarullo, *Banking on Basel: the future of international financial regulation*, 26 (Peterson Institute, 2008).
26. Ishay Weissman, "Estimation of parameters and large quantiles based on the k largest observations," In: *Journal of the American Statistical Association* **73** (27), 812–815 (1978).
27. Julien Worms and Rym Worms, "Estimation of second order parameters using probability weighted moments," In: *ESAIM: Probability and Statistics* **16** (28), 97–113 (2012).



Existence of almost automorphic solution in distribution for a class of stochastic integro-differential equation driven by Lévy noise

Mamadou Moustapha Mbaye¹ · Solym Mawaki Manou-Abi^{2,3}

Received: 17 April 2022 / Accepted: 21 January 2023

© The Author(s), under exclusive licence to The Forum D'Analyses 2023

Abstract

We investigate a class of stochastic integro-differential equations driven by Lévy noise. Under some appropriate assumptions, we establish the existence of a square-mean almost automorphic solution in distribution. Particularly, based on Schauder's fixed point theorem, the existence of square-mean almost automorphic mild solution in distribution is obtained by using the condition which is weaker than Lipschitz conditions. We provide an example to illustrate our results.

Keywords Almost automorphic solution · Stochastic processes · Stochastic evolution equations · Lévy noise

Mathematics Subject Classification 34C27 · 34K14 · 34K30 · 34K50 · 35B15 · 35K55 · 43A60 · 60G20

Communicated by S Ponnusamy.

✉ Mamadou Moustapha Mbaye
mamadoumoustapha3.mbaye@ucad.edu.sn

Solym Mawaki Manou-Abi
solym-mawaki.manou-abi@umontpellier.fr; solym.manou-abi@univ-mayotte.fr

¹ Département de Mathématiques, Faculté des Sciences et Technique, Université Cheikh Anta Diop, BP-5005, Dakar-Fann, Senegal

² Institut Montpellierain Alexander Grothendieck, UM CNRS 5149, Université de Montpellier, Montpellier, France

³ Département Sciences et Technologies, CUFR de Mayotte, 3 Route Nationale, 97660 Dembéné, France

1 Introduction

The aim of this work is to study the existence of a square-mean almost automorphic mild solutions in distribution to the following class of nonlinear stochastic integro-differential equations driven by Lévy noise in a separable Hilbert space \mathbb{H}

$$\begin{aligned}
 x'(t) = & Ax(t) + g(t, x(t)) + \int_{-\infty}^t B_1(t-s)f(s, x(s))ds \\
 & + \int_{-\infty}^t B_2(t-s)h(s, x(s))dW(s) \\
 & + \int_{-\infty}^t B_2(t-s) \int_{|y|_V < 1} F(s, x((s-), y))\tilde{N}(ds, dy) \\
 & + \int_{-\infty}^t B_2(t-s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy) \quad \text{for all } t \in \mathbb{R},
 \end{aligned}
 \tag{1.1}$$

where $A : D(A) \subset H$ is the infinitesimal generator of a C_0 -semigroup $(T(t))_{t \geq 0}$, B_1 and B_2 are convolution-type kernels in $L^1(0, \infty)$ and $L^2(0, \infty)$ respectively. $g, f : \mathbb{R} \times L^2(P, \mathbb{H}) \rightarrow L^2(P, \mathbb{H})$ $h : \mathbb{R} \times L^2(P, \mathbb{H}) \rightarrow L(V, L^2(P, \mathbb{H}))$ $F, G : \mathbb{R} \times L^2(P, \mathbb{H}) \times V \rightarrow L^2(P, \mathbb{H})$; W and N are the Lévy-Itô decomposition components of the two-sided Lévy process L (with assumptions stated in Sect. 2.).

Throughout this work, we assume $(\mathbb{H}, \| \cdot \|)$ and $(V, | \cdot |)$ are real separable Hilbert spaces. We denote by $L(V, \mathbb{H})$ the family of bounded linear operators from V to \mathbb{H} and $L^2(P, \mathbb{H})$ is the space of all \mathbb{H} -valued random variables x such that

$$\mathbb{E}\|x\|^2 = \int_{\Omega} \|x\|^2 dP < + \infty.$$

The concept of almost automorphic is a natural generalization of the almost periodicity that was introduced by Bochner [6]. The basic aspects of the theory of almost automorphic functions can be found for instance into the book [23].

In recent years, the study of almost periodic or almost automorphic solutions to some stochastic differential equations have been considerably investigated in several publications [2–5, 7–11, 13, 26] because of its significance and applications in physics, mechanics and mathematical biology. The concept of square-mean almost automorphic stochastic processes was introduced by Fu and Liu [14]. As indicated in [15, 17], it appears that almost periodicity or automorphy in distribution sense is a more appropriate concept relatively to solutions of stochastic differential equations. Recently, the concept of Poisson square-mean almost automorphy was introduced by Liu and Sun [18] to deal with some stochastic evolution equations driven by Lévy noise. For the almost automorphy in distribution, its various extensions in distribution sense and the applications in stochastic differential equations, one can see [12, 16, 20, 28] for more details.

One should point out that other slightly different versions of Eq. (1.1) have been considered in the literature. In particular, under the Lipschitz conditions on g, f and h , Bezandry [3], Xia [27] and Mbaye [19] investigated the existence and uniqueness of

the solution of Eq. (1.1) in the case when $F = G = 0$ based on Banach fixed point theorem. In this work, based on Schauder's fixed point theorem, we establish that the Eq. (1.1) has at least one almost automorphic in distribution mild solution, under some conditions, notably the conditions (H.1), (H.2) and (H.4) (see Sect. 3) but with weaker than Lipschitz conditions.

The rest of this work is organized as follows. In Sect. 2, we make a recalling on Lévy process. In Sect. 3, we review some concepts and basic properties on almost automorphic and Poisson almost automorphic processes. In Sect. 4, by Schauder's fixed point theorem, we prove the existence of an square-mean almost automorphic mild solution in distribution of Eq. (1.1). In Sect. 5, we provide an example to illustrate our results.

2 Lévy process

Definition 2.1 [18] A V -valued stochastic process $L = (L(t), t \geq 0)$ is called Lévy process if:

- (1) $L(0) = 0$ almost surely;
- (2) L has independent and stationary increments;
- (3) L is stochastically continuous, i.e. for all $\epsilon > 0$ and for all $s > 0$

$$\lim_{t \rightarrow s} P(|L(t) - L(s)| > \epsilon) = 0.$$

For a given Lévy process L , the associated jump process $\Delta L = (\Delta L(t), t \geq 0)$ is given by $\Delta L(t) = L(t) - L(t-)$ for each $t \geq 0$. For any Borel set B in $V - \{0\}$, define the random counting measure

$$N(t, B)(\omega) := \#\left\{0 \leq s \leq t : \Delta L(s)(\omega) \in B\right\} := \sum_{0 \leq s \leq t} \chi_B(\Delta L(s)(\omega))$$

with $L(t-) = \lim_{t \nearrow s} L(s)$ and χ_B being the indicator function for any Borel set B in $V - \{0\}$. We define $\nu(\cdot) = \mathbb{E}(N(1, \cdot))$ and call it the intensity measure associated with L . We say that a Borel B in $V - \{0\}$ is bounded below if $0 \notin \bar{B}$, where \bar{B} is the closure of B . If B is bounded below, then $N(t, B) < \infty$ almost surely for all $t \geq 0$ and $(N(t, B), t \geq 0)$ is called Poisson random measure with intensity $\nu(B)$. So N is called Poisson random measure. For each $t \geq 0$ and B bounded below, we define the compensated Poisson random measure by

$$\tilde{N}(t, B) = N(t, B) - t\nu(B).$$

Proposition 2.1 [1, 24] Let L be the V -valued Lévy process. Then there exist $a \in V$, V -valued Wiener process W with covariance operator Q , and an independent Poisson random measure on $\mathbb{R}^+ \times (V - \{0\})$ such that for each $t \geq 0$

$$L(t) = at + W(t) + \int_{|x| < 1} x \tilde{N}(t, dx) + \int_{|x| \geq 1} x N(t, dx),$$

where the Poisson random measure N has the intensity measure ν satisfying

$$\int_V (|x|^2 \wedge 1) \nu(dx) < \infty \quad (2.1)$$

and \tilde{N} is the compensated Poisson random measure of N .

Remark 2.1 By (2.1), it follows that $\int_{|x| \geq 1} \nu(dx) < \infty$. For convenience, we denote

$$b := \int_{|x| \geq 1} \nu(dx).$$

In the sequel, Wiener processes we consider are Q -Wiener processes, for simplicity, we assume that the covariance operator Q of W is of trace class, i.e. $\text{Tr}Q < \infty$, see [25] for more details. For our purpose, we need to consider a two-sided L Lévy processes defined on $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t \in \mathbb{R}})$. It is worth mentioning that L can be obtained as follows: Let $L_1(t)$ and $L_2(t)$, $t \geq 0$ be two independent and identically distributed Lévy processes. Let

$$L(t) = \begin{cases} L_1(t) & \text{for } t \geq 0, \\ -L_2(-t) & \text{for } t < 0. \end{cases}$$

Then L is a two-sided Lévy process defined on the filtered probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t \in \mathbb{R}})$. The stochastic process $\tilde{L} = (\tilde{L}(t), t \in \mathbb{R})$ given by $\tilde{L}(t) := L(t+s) - L(s)$ for some $s \in \mathbb{R}$ is also a two-sided Lévy process which shares the same law as L . For more details about the Lévy process, we refer to [1, 18, 24].

3 Square-mean almost automorphic process

In this section, we recall the concepts of square-mean almost automorphic process and there basic properties.

Definition 3.1 Let $x : \mathbb{R} \rightarrow L^2(P, \mathbb{H})$ be a stochastic process.

(1) x is said to be stochastically bounded if there exists $M > 0$ such that

$$\mathbb{E}\|x(t)\|^2 \leq M \quad \text{for all } t \in \mathbb{R}.$$

(2) x is said to be stochastically continuous if

$$\lim_{t \rightarrow s} \mathbb{E}\|x(t) - x(s)\|^2 = 0 \quad \text{for all } s \in \mathbb{R}.$$

Denote by $SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ the space of all the stochastically bounded and continuous processes. Clearly, the space $SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ is a Banach space equipped with the following norm

$$\|x\|_{\infty} = \sup_{t \in \mathbb{R}} (\mathbb{E}\|x(t)\|^2)^{\frac{1}{2}}.$$

Definition 3.2 [18] Let $J : \mathbb{R} \times V \rightarrow L^2(P, \mathbb{H})$ be a stochastic process.

(1) J is said to be Poisson stochastically bounded if there exists $M > 0$ such that

$$\int_V \mathbb{E}\|J(t, x)\|^2 \nu(dx) \leq M \quad \text{for all } t \in \mathbb{R}.$$

(2) J is said to be Poisson stochastically continuous if

$$\lim_{t \rightarrow s} \int_V \mathbb{E}\|J(t, x) - J(s, x)\|^2 \nu(dx) = 0 \quad \text{for all } s \in \mathbb{R}.$$

Denote by $PSBC(\mathbb{R} \times V, L^2(P, \mathbb{H}))$ the space of all stochastically bounded and continuous processes.

Definition 3.3 [14] Let $x : \mathbb{R} \rightarrow L^2(P, \mathbb{H})$ be a continuous stochastic process. x is said to be square-mean almost automorphic process if for every sequence of real numbers $(t'_n)_n$ we can extract a subsequence $(t_n)_n$ such that, for some stochastic process $y : \mathbb{R} \rightarrow L^2(P, \mathbb{H})$, we have

$$\lim_{n \rightarrow +\infty} \mathbb{E}\|x(t + t_n) - y(t)\|^2 = 0 \quad \text{for all } t \in \mathbb{R}$$

and

$$\lim_{n \rightarrow +\infty} \mathbb{E}\|y(t - t_n) - x(t)\|^2 = 0 \quad \text{for all } t \in \mathbb{R}.$$

We denote the space of all such stochastic processes by $SAA(\mathbb{R}, L^2(P, \mathbb{H}))$.

Theorem 3.1 [14] $SAA(\mathbb{R}, L^2(P, \mathbb{H}))$ equipped with the norm $\|\cdot\|_{\infty}$ is a Banach space.

Definition 3.4 [18] Let $D : \mathbb{R} \times V \rightarrow L^2(P, \mathbb{H})$ be a stochastic process. D is said to be Poisson square-mean almost automorphic process in $t \in \mathbb{R}$ if D is Poisson continuous and for every sequence of real numbers $(t'_n)_n$ we can extract a subsequence $(t_n)_n$ such that, for some stochastic process $\tilde{D} : \mathbb{R} \times V \rightarrow L^2(P, \mathbb{H})$ with $\int_V \mathbb{E}\|\tilde{D}(t, x)\|^2 \nu(dx) < \infty$ such that

$$\lim_{n \rightarrow +\infty} \int_V \mathbb{E}\|D(t + t_n, x) - \tilde{D}(t, x)\|^2 \nu(dx) = 0 \quad \text{for all } t \in \mathbb{R}$$

and

$$\lim_{n \rightarrow +\infty} \int_V \mathbb{E}\|\tilde{D}(t - t_n, x) - D(t, x)\|^2 \nu(dx) = 0 \quad \text{for all } t \in \mathbb{R}.$$

We denote the space of all such stochastic processes by $PSAA(\mathbb{R} \times V, L^2(P, \mathbb{H}))$.

Definition 3.5 [18] Let $F : \mathbb{R} \times L^2(P, \mathbb{H}) \times V \rightarrow L^2(P, H)$ be stochastic process. F is said be Poisson square-mean almost automorphic process in $t \in \mathbb{R}$ for each $Y \in L^2(P, \mathbb{H})$ if F is Poisson continuous and for every sequence of real numbers $(t'_n)_n$ we can extract a subsequence $(t_n)_n$ such that, for some stochastic process $\tilde{F} : \mathbb{R} \times L^2(P, \mathbb{H}) \times V \rightarrow L^2(P, \mathbb{H})$ with $\int_V \mathbb{E}\|\tilde{F}(t, Y, x)\|^2 \nu(dx) < \infty$ such that

$$\lim_{n \rightarrow +\infty} \int_V \mathbb{E}\|F(t + t_n, Y, x) - \tilde{F}(t, Y, x)\|^2 \nu(dx) = 0 \quad \text{for all } t \in \mathbb{R}$$

and

$$\lim_{n \rightarrow +\infty} \int_V \mathbb{E}\|\tilde{F}(t - t_n, Y, x) - F(t, Y, x)\|^2 \nu(dx) = 0 \quad \text{for all } t \in \mathbb{R}.$$

We denote the space off all such stochastic processes by $PSAA(\mathbb{R} \times L^2(P, \mathbb{H}) \times V, L^2(P, \mathbb{H}))$.

Let (\mathbb{Y}, d) be a separable, complete metric space and $\mathcal{P}(\mathbb{Y})$ be the space of Borel probability measures on \mathbb{Y} . For $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{Y})$, we define

$$d_{BL}(\mu_1, \mu_2) = \sup_{\|g\|_{BL} \leq 1} \left| \int_{\mathbb{Y}} g d(\mu_1 - \mu_2) \right|, \tag{3.1}$$

where g are Lipschitz continuous functions on \mathbb{Y} with the norm

$$\|g\|_L = \sup \left\{ \frac{|g(k) - g(l)|}{d(k, l)}; k, l \in \mathbb{Y}, k \neq l \right\}$$

$$\|g\|_{BL} = \max\{\|g\|_{\infty}, \|g\|_L\}, \quad \|g\|_{\infty} := \sup_{k \in \mathbb{Y}} |g(k)| < \infty.$$

It is known that d_{BL} is a complete metric on $\mathcal{P}(\mathbb{Y})$ which generates the weak topology [21].

Definition 3.6 [18] An \mathbb{H} -valued stochastic process $Y(t)$ is said to be almost automorphic in distribution if its law $\mu(t)$ is a $\mathcal{P}(\mathbb{H})$ -valued almost automorphic mapping, i.e. for every sequence of real numbers $(s'_n)_n$, there exist a subsequence $(s_n)_n$ and a $\mathcal{P}(\mathbb{H})$ -valued mapping $\tilde{\mu}(t)$ such that

$$\lim_{n \rightarrow \infty} d_{BL}(\mu(t + s_n), \tilde{\mu}(t)) = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} d_{BL}(\tilde{\mu}(t - s_n), \mu(t)) = 0$$

hold for each $t \in \mathbb{R}$.

Remark 3.1 A square-mean almost automorphic stochastic process is necessarily an almost automorphic in distribution one; but the converse is not true. One can see [16] for more details.

To study the existence of mild solutions to the stochastic evolution equations (1.1). we will need the following assumptions,

(H.1) The semigroup $T(t)$ is compact for $t > 0$ and is exponentially stable, i.e., there exists constants $K, \omega > 0$ such that

$$\|T(t)\| \leq Ke^{-\omega t} \quad \text{for all } t \geq 0. \tag{3.2}$$

(H.2) The processes f, g and h are uniformly continuous on any bounded subset $D \subset L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$. F, G are uniformly continuous on any bounded subset $D \subset L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$ and $x \in V$. For each bounded subset $D \subset L^2(P, \mathbb{H})$, $g(\mathbb{R}, D), f(\mathbb{R}, D), h(\mathbb{R}, D)$ are bounded and $F(\mathbb{R}, D, V)$ and $G(\mathbb{R}, D, V)$ are Poisson stochastically bounded. Moreover we suppose that there exists $r > 0$ such that

$$\Delta_r \leq \frac{\omega^2 r}{20\theta K^2} \tag{3.3}$$

where

$$\Delta_r = \max \left\{ \begin{aligned} &\sup_{t \in \mathbb{R}, \|u\|_{L^2} \leq r} \|f(t, u)\|_{L^2}, \sup_{t \in \mathbb{R}, \|u\|_{L^2} \leq r} \|g(t, u)\|_{L^2}, \sup_{t \in \mathbb{R}, \|u\|_{L^2} \leq r} \|h(t, u)\|_{L^2}, \\ &\sup_{t \in \mathbb{R}, \|u\|_{L^2} \leq r} \int_{|y|_V < 1} \|F(t, u, x)\|_{L^2}^2 \nu(dx), \\ &\sup_{t \in \mathbb{R}, \|u\|_{L^2} \leq r} \int_{|y|_V \geq 1} \|G(t, u, x)\|_{L^2}^2 \nu(dx) \end{aligned} \right\}$$

and

$$\theta = \max \left(1, \|B_1\|_{L^1(0, \infty)}^2, 4\|B_2\|_{L^2(0, \infty)}^2, 2b\|B_2\|_{L^1(0, \infty)} \right).$$

(H.3) we suppose that there exist measurable functions $m_g, m_f, m_h, m_F, m_G : \mathbb{R} \rightarrow [0, \infty)$ such that

$$\mathbb{E} \|g(t, Y) - g(t, Z)\|^2 \leq m_g(t) \cdot \mathbb{E} \|Y - Z\|^2, \tag{3.4}$$

$$\mathbb{E} \|f(t, Y) - f(t, Z)\|^2 \leq m_f(t) \cdot \mathbb{E} \|Y - Z\|^2, \tag{3.5}$$

$$\mathbb{E} \| (h(t, Y) - h(t, Z)) \mathcal{Q}^{\frac{1}{2}} \|_{L(V, L^2(P, \mathbb{H}))}^2 \leq m_h(t) \cdot \mathbb{E} \|Y - Z\|^2, \tag{3.6}$$

$$\int_{|x|_V < 1} \mathbb{E} \|F(t, Y, x) - F(t, Z, x)\|^2 \nu(dx) \leq m_F(t) \cdot \mathbb{E} \|Y - Z\|^2, \tag{3.7}$$

$$\int_{|x|_V \geq 1} \mathbb{E} \|G(t, Y, x) - G(t, Z, x)\|^2 \nu(dx) \leq m_G(t) \cdot \mathbb{E} \|Y - Z\|^2, \tag{3.8}$$

for all $t \in \mathbb{R}$ and for any $Y, Z \in L^2(P, \mathbb{H})$.

(H.4) If $(u_n)_{n \in \mathbb{N}} \subset SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ is uniformly bounded and uniformly convergent upon every compact subset of \mathbb{R} , then $g(\cdot, u_n(\cdot))$, $f(\cdot, u_n(\cdot))$, $h(\cdot, u_n(\cdot))$, and $F(\cdot, u_n(\cdot), \cdot)$, $G(\cdot, u_n(\cdot), \cdot)$ are relatively compact in $SBC(\mathbb{R}, L^2(P, \mathbb{H}))$, $PSBC(\mathbb{R} \times V, L^2(P, \mathbb{H}))$, respectively.

Definition 3.7 An \mathcal{F}_t -progressively measurable process $\{x(t)\}_{t \in \mathbb{R}}$ is called a mild solution on \mathbb{R} of Eq. (1.1) if it satisfies the corresponding stochastic integral equation

$$\begin{aligned}
 x(t) &= T(t-a)x(a) + \int_a^t T(t-s)g(s, x(s))ds \\
 &+ \int_a^t T(t-\sigma) \int_a^\sigma B_1(\sigma-s)f(s, x(s))dsd\sigma \\
 &+ \int_a^t T(t-\sigma) \int_a^\sigma B_2(\sigma-s)h(s, x(s))dW(s)d\sigma \\
 &+ \int_a^t T(t-\sigma) \int_a^\sigma B_2(\sigma-s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
 &+ \int_a^t T(t-\sigma) \int_a^\sigma B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma
 \end{aligned} \tag{3.9}$$

for all $t \geq a$.

Remark 3.2 If we let $a \rightarrow -\infty$ in the stochastic integral equation (3.9), by the exponential dissipation condition of $(T(t))_{t \geq 0}$, then we obtain the stochastic process $x : \mathbb{R} \rightarrow L^2(\Omega, \mathbb{H})$ is a mild solution of the Eq. (3.9) if and only if x satisfies the stochastic integral equation

$$\begin{aligned}
 x(t) &= \int_{-\infty}^t T(t-s)g(s, x(s))ds + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^\sigma B_1(\sigma-s)f(s, x(s))dsd\sigma \\
 &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^\sigma B_2(\sigma-s)h(s, x(s))dW(s)d\sigma \\
 &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^\sigma B_2(\sigma-s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
 &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^\sigma B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma.
 \end{aligned} \tag{3.10}$$

4 Square-mean almost automorphic solutions

This section is devoted to the existence of the square-mean almost automorphic mild solution in distribution on \mathbb{R} of Eq. (1.1).

Define the following integral operator,

$$\begin{aligned}
(\Lambda x)(t) &= \int_{-\infty}^t T(t-s)g(s, x(s))ds + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)f(s, x(s))dsd\sigma \\
&+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)h(s, x(s))dW(s)d\sigma \\
&+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
&+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma.
\end{aligned}$$

We have

Lemma 4.1 *Suppose that the processes f, g and h are uniformly continuous on any bounded subset $D \subset L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$. F, G are uniformly continuous on any bounded subset $D \subset L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$ and $x \in V$. For each bounded subset $D \subset L^2(P, \mathbb{H})$, $g(\mathbb{R}, D)$, $f(\mathbb{R}, D)$, $h(\mathbb{R}, D)$ are bounded and $F(\mathbb{R}, D, V)$ and $G(\mathbb{R}, D, V)$ are Poisson stochastically bounded. If the semigroup $T(t)$ verifies (3.2) in assumption (H.1) then the mapping $\Lambda : SBC(\mathbb{R}, L^2(P, \mathbb{H})) \rightarrow SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ is well-defined and continuous.*

Proof It is easy to see that S is well-defined. To complete the proof it remains to see that Λ is continuous. Consider an arbitrary sequence of functions $u_n \in SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ that converges uniformly to some $u \in SBC(\mathbb{R}, L^2(P, \mathbb{H}))$, that is, $\|u_n - u\|_{\infty} \rightarrow 0$ as $n \rightarrow \infty$. There exists a bounded subset D of $L^2(P, \mathbb{H})$ such that $u_n(t), u(t) \in D$ for each $t \in \mathbb{R}$ and $n = 1, 2, \dots$. By assumptions, given $\epsilon > 0$, there exist $\delta > 0$ and $N > 0$ such that $\mathbb{E}\|u_n(t) - u(t)\|^2 < \delta$ imply that

$$\mathbb{E} \|g(t, u_n(t)) - g(t, u(t))\|^2 < \frac{\omega^2 \epsilon}{25K^2},$$

$$\mathbb{E} \|f(t, u_n(t)) - f(t, u(t))\|^2 < \frac{\omega^2 \epsilon}{25K^2(\|B_1\|_{L^1(0, \infty)}^2 + 1)},$$

$$\mathbb{E} \|h(t, u_n(t)) - h(t, u(t))\|_{Q^{\frac{1}{2}}}^2 < \frac{\omega^2 \epsilon}{25K^2(\|B_2\|_{L^2(0, \infty)}^2 + 1)},$$

$$\int_{|x|_V < 1} \mathbb{E}\|F(t, u_n(t), x) - F(t, u(t), x)\|^2 \nu(dx) < \frac{\omega^2 \epsilon}{25K^2(\|B_2\|_{L^2(0, \infty)}^2 + 1)},$$

$$\text{and } \int_{|x|_V < 1} \mathbb{E}\|G(t, u_n(t), x) - G(t, u(t), x)\|^2 \nu(dx) < \frac{\omega^2 \epsilon}{50K^2(\|B_2\|_{L^2(0, \infty)}^2 + b\|B_2\|_{L^1(0, \infty)}^2 + 1)}$$

for all $t \in \mathbb{R}$ and $n \geq N$. Hence

$$\begin{aligned}
\mathbb{E}\|(\Lambda u_n)(t) - (\Lambda u)(t)\|^2 &= \mathbb{E}\left\| \int_{-\infty}^t T(t-s) \left(g(s, u_n(s)) - g(s, u(s)) \right) ds \right. \\
&\quad + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s) \left(f(s, u_n(s)) - f(s, u(s)) \right) ds d\sigma \\
&\quad + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \left(h(s, u_n(s)) - h(s, u(s)) \right) dW(s) d\sigma \\
&\quad + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \\
&\quad \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) \tilde{N}(ds, dy) d\sigma \\
&\quad + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} \\
&\quad \left(G(s, u_n(s-), y) - G(s, u(s-), y) \right) N(ds, dy) d\sigma \left. \right\|^2 \\
&\leq 5\mathbb{E}\left\| \int_{-\infty}^t T(t-s) \left(g(s, u_n(s)) - g(s, u(s)) \right) ds \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s) \left(f(s, u_n(s)) - f(s, u(s)) \right) ds d\sigma \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \left(h(s, u_n(s)) - h(s, u(s)) \right) dW(s) d\sigma \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) \right. \\
&\quad \left. \tilde{N}(ds, dy) d\sigma \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} \left(G(s, u_n(s-), y) - G(s, u(s-), y) \right) \right. \\
&\quad \left. N(ds, dy) d\sigma \right\|^2 \\
&\leq 5(I_1 + I_2 + I_3 + I_4 + I_5).
\end{aligned}$$

Using Cauchy–Schwartz’s inequality, we get

$$I_1 < \frac{\epsilon}{25} \quad I_2 < \frac{\epsilon}{25}.$$

For I_3 , using Cauchy–Schwartz’s inequality and Ito’s isometry property, we obtain

$$I_3 \leq \frac{K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \mathbb{E} \left\| B_2(\sigma - s) \left(h(s, u_n(s)) - h(s, u(s)) \right) Q^{\frac{1}{2}} \right\|^2 ds d\sigma < \frac{\epsilon}{25}.$$

As to I_4 and I_5 , by Cauchy–Schwartz’s inequality and the properties of the integral for the Poisson random measure, we have

$$\begin{aligned} I_4 &= \mathbb{E} \left\| \int_{-\infty}^t T(t - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\ &\leq \frac{K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \mathbb{E} \left\| \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) \tilde{N}(ds, dy) \right\|^2 d\sigma \\ &\leq \frac{K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma - s)\|^2 \int_{|y|_V < 1} \mathbb{E} \left\| F(s, u_n(s-), y) - F(s, u(s-), y) \right\|^2 v(dy) d\sigma \\ &< \frac{\epsilon}{25}. \end{aligned}$$

And

$$\begin{aligned} I_5 &= \mathbb{E} \left\| \int_{-\infty}^t T(t - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) N(ds, dy) d\sigma \right\|^2 \\ &\leq 2 \mathbb{E} \left\| \int_{-\infty}^t T(t - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\ &\quad + 2 \mathbb{E} \left\| \int_{-\infty}^t T(t - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) v(dy) d\sigma \right\|^2 \\ &\leq \frac{2K^2}{\omega^2} \|B_2\|_{L^2(0, \infty)}^2 \left(\frac{\omega^2 \epsilon}{50K^2 (\|B_2\|_{L^2(0, \infty)}^2 + b \|B_2\|_{L^1(0, \infty)}^2 + 1)} \right) \\ &\quad + 2 \frac{K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \mathbb{E} \left\| \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} \left(F(s, u_n(s-), y) - F(s, u(s-), y) \right) v(dy) ds \right\|^2 d\sigma \\ &\leq \frac{2K^2}{\omega^2} \|B_2\|_{L^2(0, \infty)}^2 \left(\frac{\omega^2 \epsilon}{50K^2 (\|B_2\|_{L^2(0, \infty)}^2 + b \|B_2\|_{L^1(0, \infty)}^2 + 1)} \right) \\ &\quad + 2 \frac{K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \left(\int_{-\infty}^{\sigma} \|B_2(\sigma - s)\| ds \int_{|y|_V \geq 1} v(dy) \right. \\ &\quad \cdot \left. \int_{-\infty}^{\sigma} \|B_2(\sigma - s)\| \int_{|y|_V \geq 1} \mathbb{E} \left\| F(s, u_n(s-), y) - F(s, u(s-), y) \right\|^2 v(dy) ds \right) d\sigma \\ &\leq \frac{2K^2}{\omega^2} \|B_2\|_{L^2(0, \infty)}^2 \left(\frac{\omega^2 \epsilon}{50K^2 (\|B_2\|_{L^2(0, \infty)}^2 + b \|B_2\|_{L^1(0, \infty)}^2 + 1)} \right) \\ &\quad + 2b \frac{K^2}{\omega} \|B_2\|_{L^1(0, \infty)}^2 \left(\frac{\omega^2 \epsilon}{50K^2 (\|B_2\|_{L^2(0, \infty)}^2 + b \|B_2\|_{L^1(0, \infty)}^2 + 1)} \right) \\ &< \frac{\epsilon}{25}. \end{aligned}$$

Thus, by combining $I_1 - I_5$, it follows that for each $t \in \mathbb{R}$ and $n \geq N$

$$\mathbb{E}\|(\Lambda u_n)(t) - (\Lambda u)(t)\|^2 < \epsilon.$$

This implies that Λ is continuous. The proof is complete. \square

Theorem 4.1 *Assume that assumptions (H.1)–(H.4) hold and*

- (1) $g, f \in SAA(\mathbb{R} \times L^2(P, \mathbb{H}), L^2(P, \mathbb{H}))$,
- (2) $h \in SAA(\mathbb{R} \times L^2(P, \mathbb{H}), L(V, L^2(P, \mathbb{H})))$,
- (3) $F \in PSAA(\mathbb{R} \times L^2(P, \mathbb{H}) \times V, L^2(P, \mathbb{H}))$
- (4) $G \in PSAA(\mathbb{R} \times L^2(P, \mathbb{H}) \times V, L^2(P, \mathbb{H}))$.

then Eq. (1.1) has at least one almost automorphic in distribution mild solution on \mathbb{R} provided that

$$\begin{aligned} L_g &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t e^{-\omega(t-s)} m_g(s) ds < \infty, & L_f &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_1(t-s)\| m_f(s) ds < \infty, \\ L_h &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 m_h(s) ds < \infty, & L_F &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 m_F(s) ds < \infty, \\ L_G &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 m_G(s) ds < \infty \end{aligned}$$

and

$$\vartheta := 10 \frac{K^2}{\omega^2} \left[\omega L_g + L_f \|B_1\|_{L^1(0, \infty)} + L_h + L_F + 2 \left(1 + b \|B_2\|_{L^1(0, \infty)} \right) L_G \right] < 1. \quad (4.1)$$

Proof Let $B = \{u \in SBC(\mathbb{R}, L^2(P, \mathbb{H})) : \|u\|_{\infty}^2 \leq r\}$. By Lemma 4.1 and (3.3), it follows that B is a convex and closed set satisfying $\Lambda B \subset B$. To complete the proof, we have to prove the following statements:

- (a) That $V = \{\Lambda u(t) : u \in B\}$ is a relatively compact subset of $L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$;
- (b) That $U = \{\Lambda u : u \in B\} \subset SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ is equi-continuous.
- (c) The mild solution is almost automorphic in distribution.

To prove (a), fix $t \in \mathbb{R}$ and consider an arbitrary $\varepsilon > 0$. Then

$$\begin{aligned}
(\Lambda_\epsilon x)(t) &= \int_{-\infty}^{t-\epsilon} T(t-s)g(s, x(s))ds + \int_{-\infty}^{t-\epsilon} T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)f(s, x(s))dsd\sigma \\
&\quad + \int_{-\infty}^{t-\epsilon} T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)h(s, x(s))dW(s)d\sigma \\
&\quad + \int_{-\infty}^{t-\epsilon} T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
&\quad + \int_{-\infty}^{t-\epsilon} T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \\
&= T(\epsilon) \int_{-\infty}^{t-\epsilon} T(t-\epsilon-s)g(s, x(s))ds + \int_{-\infty}^{t-\epsilon} T(t-\epsilon-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)f(s, x(s))dsd\sigma \\
&\quad + \int_{-\infty}^{t-\epsilon} T(t-\epsilon-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)h(s, x(s))dW(s)d\sigma \\
&\quad + \int_{-\infty}^{t-\epsilon} T(t-\epsilon-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
&\quad + \int_{-\infty}^{t-\epsilon} T(t-\epsilon-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \\
&= T(\epsilon)(\Lambda x)(t-\epsilon).
\end{aligned}$$

and hence $V_\epsilon := \{(\Lambda_\epsilon)x(t) : x \in B\}$ is relatively compact in $L^2(P, \mathbb{H})$ as the semi-group family $T(\epsilon)$ is compact by assumption.

Now

$$\begin{aligned}
\mathbb{E}\|(\Lambda u)(t) - (\Lambda_\epsilon u)(t)\|^2 &= \mathbb{E}\left\| \int_{t-\epsilon}^t T(t-s)g(s, u(s))ds \right. \\
&\quad + \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)f(s, u(s))dsd\sigma \\
&\quad + \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)h(s, u(s))dW(s)d\sigma \\
&\quad + \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} F(s, u(s-), y)\tilde{N}(ds, dy)d\sigma \\
&\quad \left. + \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, u(s-), y)N(ds, dy)d\sigma \right\|^2 \\
&\leq 5\mathbb{E}\left\| \int_{t-\epsilon}^t T(t-s)g(s, u(s))ds \right\|^2 + 5\mathbb{E}\left\| \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)f(s, u(s))dsd\sigma \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)h(s, u(s))dW(s)d\sigma \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} F(s, u(s-), y)\tilde{N}(ds, dy)d\sigma \right\|^2 \\
&\quad + 5\mathbb{E}\left\| \int_{t-\epsilon}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} G(s, u(s-), y)N(ds, dy)d\sigma \right\|^2 \\
&\leq 5\left[\Delta_r K^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 + \Delta_r K^2 \|B_1\|_{L^1(0, \infty)}^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 \right]
\end{aligned}$$

$$\begin{aligned}
 & + \Delta_r K^2 \|B_2\|_{L^2(0,\infty)}^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 + \Delta_r K^2 \|B_2\|_{L^2(0,\infty)}^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 \\
 & + \left(2\Delta_r K^2 \|B_2\|_{L^2(0,\infty)}^2 + 2b\Delta_r K^2 \|B_2\|_{L^1(0,\infty)}^2 \right) \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 \\
 & \leq 5\Delta_r K^2 \left[1 + \|B_1\|_{L^1(0,\infty)}^2 + 4\|B_2\|_{L^2(0,\infty)}^2 + 2b\|B_2\|_{L^1(0,\infty)}^2 \right] \epsilon^2
 \end{aligned}$$

from which it follows that $V = \{\Lambda u(t) : u \in B\}$ is a relatively compact subset of $L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$.

We now show that (b) holds. Let $u \in B$ and $t_1, t_2 \in \mathbb{R}$ such that $t_1 < t_2$. Similar computation as that in Lemma 4.1, we have

$$\begin{aligned}
 \mathbb{E} \|(\Lambda u)(t_1) - (\Lambda u)(t_2)\|^2 &= \mathbb{E} \left\| \int_{-\infty}^{t_2} T(t_2 - s)g(s, x(s))ds \right. \\
 &+ \int_{-\infty}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_1(\sigma - s)f(s, x(s))dsd\sigma \\
 &+ \int_{-\infty}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s)h(s, x(s))dW(s)d\sigma \\
 &+ \int_{-\infty}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
 &+ \int_{-\infty}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \\
 &- \left(\int_{-\infty}^{t_1} T(t_1 - s)g(s, x(s))ds + \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_1(\sigma - s)f(s, x(s))dsd\sigma \right. \\
 &+ \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s)h(s, x(s))dW(s)d\sigma \\
 &+ \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
 &\left. + \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \right) \Big\|^2 \\
 &= \mathbb{E} \left\| \left(T(t_2 - t_1) - I \right) \left[\int_{-\infty}^{t_1} T(t_1 - s)g(s, x(s))ds + \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_1(\sigma - s)f(s, x(s))dsd\sigma \right. \right. \\
 &+ \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s)h(s, x(s))dW(s)d\sigma \\
 &+ \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
 &\left. + \int_{-\infty}^{t_1} T(t_1 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \right] \\
 &\left. + \int_{t_1}^{t_2} T(t_2 - s)g(s, x(s))ds + \int_{t_1}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_1(\sigma - s)f(s, x(s))dsd\sigma \right\}
 \end{aligned}$$

$$\begin{aligned}
 &+ \int_{t_1}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s)h(s, x(s))dW(s)d\sigma \\
 &+ \int_{t_1}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \\
 &+ \int_{t_1}^{t_2} T(2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma \\
 &\quad - s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \Big\|^2 \leq 6 \sup_{y \in V} \mathbb{E} \left\| \left(T(t_2 - t_1) - I \right) y \right\|^2 \\
 &+ 6\mathbb{E} \left\| \int_{t_1}^{t_2} T(t_2 - s)g(s, x(s))ds \right\|^2 \\
 &+ 6\mathbb{E} \left\| \int_{t_1}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_1(\sigma - s)f(s, x(s))dsd\sigma \right\|^2 \\
 &+ 6\mathbb{E} \left\| \int_{t_1}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s)h(s, x(s))dW(s)d\sigma \right\|^2 \\
 &+ 6\mathbb{E} \left\| \int_{t_1}^{t_2} T(t_2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma - s) \int_{|y|_V < 1} F(s, x(s-), y)\tilde{N}(ds, dy)d\sigma \right\|^2 \\
 &+ 6\mathbb{E} \left\| \int_{t_1}^{t_2} T(2 - \sigma) \int_{-\infty}^{\sigma} B_2(\sigma \\
 &\quad - s) \int_{|y|_V \geq 1} G(s, x(s-), y)N(ds, dy)d\sigma \right\|^2 \leq 6 \sup_{y \in V} \mathbb{E} \left\| \left(T(t_2 - t_1) - I \right) y \right\|^2 \\
 &+ 6 \left[\Delta_r K^2 \left(\int_{t_1}^{t_2} e^{-\omega(t_2 - \sigma)} d\sigma \right)^2 \right. \\
 &+ \Delta_r K^2 \|B_1\|_{L^1(0, \infty)}^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 \\
 &+ \Delta_r K^2 \|B_2\|_{L^2(0, \infty)}^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 + \Delta_r K^2 \|B_2\|_{L^2(0, \infty)}^2 \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 \\
 &+ \left(2\Delta_r K^2 \|B_2\|_{L^2(0, \infty)}^2 \right. \\
 &\quad \left. + 2b\Delta_r K^2 \|B_2\|_{L^1(0, \infty)}^2 \right) \left(\int_{t-\epsilon}^t e^{-\omega(t-\sigma)} d\sigma \right)^2 \Big] \leq 6 \sup_{y \in V} \mathbb{E} \left\| \left(T(t_2 - t_1) - I \right) y \right\|^2 \\
 &+ 6\Delta_r K^2 \left[1 + \|B_1\|_{L^1(0, \infty)}^2 + 4\|B_2\|_{L^2(0, \infty)}^2 + 2b\|B_2\|_{L^1(0, \infty)}^2 \right] \left(\int_{t_1}^{t_2} e^{-\omega(t_2 - \sigma)} d\sigma \right)^2.
 \end{aligned}$$

The right-hand side tends to 0 independently to $u \in B$ as $t_2 \rightarrow t_1$ which implies that U is right equi-continuous at t . Similarly, we can show that U is left equi-continuous at t .

Denote the closed convex hull of ΛB by $\overline{co} \Lambda B$. Since $\overline{co} \Lambda B \subset B$ and B is a closed convex, it follows that

$$\Lambda(\overline{co} \Lambda B) \subset \Lambda B \subset \overline{co} \Lambda B.$$

Further, it is not hard to see that $\left\{x(t) : x \in \overline{c\partial \Lambda B}\right\}$ is relatively compact in $L^2(P, \mathbb{H})$ for each $t \in \mathbb{R}$ and $\overline{c\partial \Lambda B} \subset SBC(\mathbb{R}, L^2(P, \mathbb{H}))$ is uniformly bounded and equi-continuous. Using Arzelà-Ascoli theorem, we deduce that the restriction of $\overline{c\partial \Lambda B}$ to any compact subset I of \mathbb{R} is relatively compact in $C(I, L^2(P, \mathbb{H}))$. Thus condition (H.4) implies that $\Lambda : \overline{c\partial \Lambda B} \rightarrow \overline{c\partial \Lambda B}$ is a compact operator. In summary, $\Lambda : \overline{c\partial \Lambda B} \rightarrow \overline{c\partial \Lambda B}$ is continuous and compact. Using the Schauder fixed point theorem, it follows that Λ has a fixed-point.

To end the proof, we have to check this the fixed-point is almost automorphic in distribution. Since g, f, h are almost automorphic and F, G are Poisson almost automorphic, then for every sequence of real numbers $(t'_n)_n$ we can extract a subsequence $(t_n)_n$ such that, for some stochastic processes $\tilde{g}, \tilde{f}, \tilde{h}, \tilde{F}, \tilde{G}$

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|g(s + t_n, X) - \tilde{g}(s)\|^2 = 0, \quad \lim_{n \rightarrow +\infty} \mathbb{E} \|\tilde{g}(s - t_n, X) - g(s, X)\|^2 = 0; \tag{4.2}$$

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|f(s + t_n, X) - \tilde{f}(s, X)\|^2 = 0, \quad \lim_{n \rightarrow +\infty} \mathbb{E} \|\tilde{f}(s - t_n, X) - f(s, X)\|^2 = 0; \tag{4.3}$$

$$\begin{aligned} \lim_{n \rightarrow +\infty} \mathbb{E} \left\| \left(h(s + t_n, X) - \tilde{h}(s, X) \right) \mathcal{Q}^\perp \right\|_{L(V, L^2(P, \mathbb{H}))}^2 &= 0, \\ \lim_{n \rightarrow +\infty} \mathbb{E} \left\| \left(\tilde{h}(s - t_n, X) - h(s, X) \right) \mathcal{Q}^\perp \right\|_{L(V, L^2(P, \mathbb{H}))}^2 &= 0; \end{aligned} \tag{4.4}$$

$$\begin{aligned} \lim_{n \rightarrow +\infty} \int_{|y|_V < 1} \mathbb{E} \|F(s + t_n, X, y) - \tilde{F}(s, X, y)\|^2 \nu(dy) &= 0, \\ \lim_{n \rightarrow +\infty} \int_{|y|_V < 1} \mathbb{E} \|\tilde{F}(s - t_n, X, y) - F(s, X, y)\|^2 \nu(dy) &= 0 \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} \lim_{n \rightarrow +\infty} \int_{|y|_V \geq 1} \mathbb{E} \|G(s + t_n, X, y) - \tilde{G}(s, X, y)\|^2 \nu(dy) &= 0, \\ \lim_{n \rightarrow +\infty} \int_{|y|_V \geq 1} \mathbb{E} \|\tilde{G}(s - t_n, X, y) - G(s, X, y)\|^2 \nu(dy) &= 0 \end{aligned} \tag{4.6}$$

hold for each $s \in \mathbb{R}$ and $X \in L^2(P, H)$.

For $t \in \mathbb{R}$, we define

$$\begin{aligned} \tilde{X}(t) &= \int_{-\infty}^t T(t-s)\tilde{g}(s, \tilde{X}(s))ds + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)\tilde{f}(s, \tilde{X}(s))dsd\sigma \\ &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)\tilde{h}(s, \tilde{X}(s))dW(s)d\sigma \\ &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \tilde{F}(s, \tilde{X}(s-), y)\tilde{N}(ds, dy)d\sigma \\ &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} \tilde{G}(s, \tilde{X}(s-), y)N(ds, dy)d\sigma. \end{aligned}$$

Let $W_n(s) = W(s + t_n) - W(t_n)$, $N_n(s, y) = N(s + t_n, y) - N(t_n, x)$ and $\tilde{N}_n(s, y) = N(s + t_n, y) - N(t_n, x)$ for each $s \in \mathbb{R}$. Then W_n is also a Q-Wiener process having the same distribution as W and N_n have the same distribution as N with compensated Poisson random measure \tilde{N}_n .

Consider the process define as follows

$$\begin{aligned} X_n(t) &= \int_{-\infty}^t T(t-s)g(s + t_n, X_n(s))ds + \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)f(s + t_n, X_n(s))dsd\sigma \\ &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)h(s + t_n, X_n(s))dW(s)d\sigma \\ &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} F(s + t_n, X_n(s-), y)\tilde{N}(ds, dy)d\sigma \\ &+ \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} G(s + t_n, X_n(s-), y)N(ds, dy)d\sigma. \end{aligned}$$

Note that $X(t + t_n)$ and X_n have the same law and since the convergence in L^2 implies convergence in distribution, then we have

$$\begin{aligned} &\mathbb{E}\|X_n(t) - \tilde{X}(t)\|^2 \\ &\leq 5\mathbb{E}\left\|\int_{-\infty}^t T(t-s)\left(g(s + t_n, X_n(s)) - \tilde{g}(s, \tilde{X}(s))\right)ds\right\|^2 \\ &+ 5\mathbb{E}\left\|\int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s)\left(f(s + t_n, X_n(s)) - \tilde{f}(s, \tilde{X}(s))\right)dsd\sigma\right\|^2 \\ &+ 5\mathbb{E}\left\|\int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s)\left(h(s + t_n, X_n(s)) - \tilde{h}(s, \tilde{X}(s))\right)dW(s)d\sigma\right\|^2 \\ &+ 5\mathbb{E}\left\|\int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \left(F(s + t_n, X_n(s-), y) - \tilde{F}(s, \tilde{X}(s-), y)\right)\tilde{N}(ds, dy)d\sigma\right\|^2 \\ &+ 5\mathbb{E}\left\|\int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V \geq 1} \left(G(s + t_n, X_n(s-), y) - \tilde{G}(s, \tilde{X}(s-), y)\right)N(ds, dy)d\sigma\right\|^2 \\ &\leq 5(J_1 + J_2 + J_3 + J_4 + J_5). \end{aligned}$$

For J_1 , we have

$$\begin{aligned}
 J_1 &= \mathbb{E} \left\| \int_{-\infty}^t T(t-s) \left(g(s+t_n, X_n(s)) - \tilde{g}(s, \tilde{X}(s)) \right) ds \right\|^2 \\
 &\leq 2\mathbb{E} \left\| \int_{-\infty}^t T(t-s) \left(g(s+t_n, X_n(s)) - g(s+t_n, \tilde{X}(s)) \right) ds \right\|^2 \\
 &\quad + 2\mathbb{E} \left\| \int_{-\infty}^t T(t-s) \left(g(s+t_n, \tilde{X}(s)) - \tilde{g}(s, \tilde{X}(s)) \right) ds \right\|^2 \\
 &\leq \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-s)} m_g(s+t_n) \mathbb{E} \left\| X_n(s) - \tilde{X}(s) \right\|^2 ds \\
 &\quad + \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-s)} \mathbb{E} \left\| g(s+t_n, \tilde{X}(s)) - \tilde{g}(s, \tilde{X}(s)) \right\|^2 ds \\
 &\leq \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-s)} m_g(s+t_n) \mathbb{E} \left\| X_n(s) - \tilde{X}(s) \right\|^2 ds + \zeta_1^n,
 \end{aligned}$$

where $\zeta_1^n := \frac{2K^2}{\omega^2} \sup_{s \in \mathbb{R}} \mathbb{E} \left\| g(s+t_n, \tilde{X}(s)) - \tilde{g}(s, \tilde{X}(s)) \right\|^2 ds$. Since $\tilde{X}(\cdot)$ is bounded in $L^2(P, H)$, it follows by (4.2), that $\zeta_1^n \rightarrow 0$ as $n \rightarrow \infty$.

For J_2 , we have

$$\begin{aligned}
 J_2 &= \mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s) \left(f(s+t_n, X_n(s)) - \tilde{f}(s, \tilde{X}(s)) \right) ds d\sigma \right\|^2 \\
 &\leq 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s) \left(f(s+t_n, X_n(s)) - f(s+t_n, \tilde{X}(s)) \right) ds d\sigma \right\|^2 \\
 &\quad + 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_1(\sigma-s) \left(f(s+t_n, \tilde{X}(s)) - \tilde{f}(s, \tilde{X}(s)) \right) ds d\sigma \right\|^2 \\
 &\leq \frac{2K^2}{\omega} \|B_1\|_{L^1(0,\infty)} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_1(\sigma-s)\| m_f(s+t_n) \mathbb{E} \left\| X_n(s) - \tilde{X}(s) \right\|^2 ds d\sigma \\
 &\quad + \frac{2K^2}{\omega} \|B_1\|_{L^1(0,\infty)} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_1(\sigma-s)\| \mathbb{E} \|f(s+t_n, \tilde{X}(s)) - \tilde{f}(s, \tilde{X}(s))\|^2 ds d\sigma \\
 &\leq \frac{2K^2}{\omega} \|B_1\|_{L^1(0,\infty)} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_1(\sigma-s)\| m_f(s+t_n) \mathbb{E} \left\| X_n(s) - \tilde{X}(s) \right\|^2 ds d\sigma + \zeta_2^n,
 \end{aligned}$$

where $\zeta_2^n := \frac{2K^2}{\omega^2} \|B_1\|_{L^1(0,\infty)}^2 \sup_{s \in \mathbb{R}} \mathbb{E} \|f(s+t_n, \tilde{X}(s)) - \tilde{f}(s, \tilde{X}(s))\|^2$. For the same reason as for ζ_1^n , $\zeta_2^n \rightarrow 0$ as $n \rightarrow \infty$.

For J_3 , using Cauchy–Schwartz’s inequality and the Ito’s isometry, we have

$$\begin{aligned}
 J_3 &= \mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \left(h(s+t_n, X_n(s)) - \tilde{h}(s, \tilde{X}(s)) \right) dW(s) d\sigma \right\|^2 \\
 &\leq 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \left(h(s+t_n, X_n(s)) - h(s+t_n, \tilde{X}(s)) \right) dW(s) d\sigma \right\|^2 \\
 &\quad + 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \left(h(s+t_n, \tilde{X}(s)) - \tilde{h}(s, \tilde{X}(s)) \right) dW(s) d\sigma \right\|^2 \\
 &\leq \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_h(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma \\
 &\quad + \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 \mathbb{E} \left\| \left(h(s+t_n, \tilde{X}(s)) - \tilde{h}(s, \tilde{X}(s)) \right) \mathcal{Q}^{\frac{1}{2}} \right\|_{L(V, L^2(P, H))}^2 ds d\sigma \\
 &\leq \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_h(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma + \zeta_3^n,
 \end{aligned}$$

where

$$\zeta_3^n := \frac{2K^2}{\omega^2} \|B_2\|_{L^2(0, \infty)}^2 \sup_{s \in \mathbb{R}} \mathbb{E} \left\| \left(h(s+t_n, \tilde{X}(s)) - \tilde{h}(s, \tilde{X}(s)) \right) \mathcal{Q}^{\frac{1}{2}} \right\|_{L(V, L^2(P, H))}^2. \quad \text{By}$$

(4.4), it follows that $\zeta_3^n \rightarrow 0$ as $n \rightarrow \infty$, like ζ_1^n .

For J_4 , using Cauchy-Schwartz's inequality and the properties of the integral for the Poisson random measure, we have

$$\begin{aligned}
 J_4 &= \mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \left(F(s+t_n, X_n(s-), y) - \tilde{F}(s, \tilde{X}(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\
 &\leq 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \left(F(s+t_n, X_n(s-), y) - F(s+t_n, \tilde{X}(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\
 &\quad + 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_V < 1} \left(F(s+t_n, \tilde{X}(s-), y) - \tilde{F}(s, \tilde{X}(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\
 &\leq \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_F(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma \\
 &\quad + \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 \int_{|y|_V < 1} \mathbb{E} \left\| F(s+t_n, \tilde{X}(s-), y) - \tilde{F}(s, \tilde{X}(s-), y) \right\|^2 v(dy) ds d\sigma \\
 &\leq \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_F(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma + \zeta_4^n,
 \end{aligned}$$

where $\zeta_4^n := \frac{2K^2}{\omega^2} \|B_2\|_{L^2(0, \infty)}^2 \sup_{s \in \mathbb{R}} \left(\int_{|y|_V < 1} \mathbb{E} \left\| F(s+t_n, \tilde{X}(s-), y) - \tilde{F}(s, \tilde{X}(s-), y) \right\|^2 v(dy) \right)$. Using (4.5), it follows that $\zeta_4^n \rightarrow 0$ as $n \rightarrow \infty$, like ζ_1^n .

For J_5 , using Cauchy-Schwartz's inequality and the properties of the integral for the Poisson random measure, we have

$$\begin{aligned}
 J_5 &= \mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, X_n(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right) N(ds, dy) d\sigma \right\|^2 \\
 &\leq 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, X_n(s-), y) - G(s+t_n, \tilde{X}(s-), y) \right) N(ds, dy) d\sigma \right\|^2 \\
 &\quad + 2\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right) N(ds, dy) d\sigma \right\|^2 \\
 &\leq 4\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, X_n(s-), y) - G(s+t_n, \tilde{X}(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\
 &\quad + 4\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, X_n(s-), y) - G(s+t_n, \tilde{X}(s-), y) \right) v(dy) ds d\sigma \right\|^2 \\
 &\quad + 4\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right) \tilde{N}(ds, dy) d\sigma \right\|^2 \\
 &\quad + 4\mathbb{E} \left\| \int_{-\infty}^t T(t-\sigma) \int_{-\infty}^{\sigma} B_2(\sigma-s) \int_{|y|_r \geq 1} \left(G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right) v(dy) ds d\sigma \right\|^2 \\
 &\leq \frac{4K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_G(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma \\
 &\quad + \frac{4K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \left(\int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| \int_{|y|_r \geq 1} v(dy) ds \right. \\
 &\quad \times \left. \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| \int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, X_n(s-), y) - G(s+t_n, \tilde{X}(s-), y) \right\|^2 v(dy) ds \right) d\sigma \\
 &\quad + \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 \int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right\|^2 v(dy) ds d\sigma \\
 &\quad + \frac{4K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \left(\int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| \int_{|y|_r \geq 1} v(dy) ds \right. \\
 &\quad \times \left. \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| \int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right\|^2 v(dy) ds \right) d\sigma \\
 &\leq \frac{4K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_G(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma \\
 &\quad + \frac{4bK^2}{\omega} \|B_2\|_{L^1(0,\infty)} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| m_G(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma \\
 &\quad + \frac{2K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 \int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right\|^2 v(dy) ds d\sigma \\
 &\quad + \frac{4bK^2}{\omega} \|B_2\|_{L^2(0,\infty)} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| \int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right\|^2 v(dy) ds d\sigma \\
 &\leq \frac{4K^2}{\omega} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\|^2 m_G(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma \\
 &\quad + \frac{4bK^2}{\omega} \|B_2\|_{L^1(0,\infty)} \int_{-\infty}^t e^{-\omega(t-\sigma)} \int_{-\infty}^{\sigma} \|B_2(\sigma-s)\| m_G(s+t_n) \mathbb{E} \|X_n(s) - \tilde{X}(s)\|^2 ds d\sigma + \zeta_5^n(t)
 \end{aligned}$$

where

$$\begin{aligned}
 \zeta_5^n &= \frac{2K^2}{\omega^2} \|B_2\|_{L^2(0,\infty)}^2 \sup_{s \in \mathbb{R}} \left(\int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right\|^2 v(dy) \right) \\
 &\quad + \frac{4bK^2}{\omega^2} \|B_2\|_{L^1(0,\infty)}^2 \sup_{s \in \mathbb{R}} \left(\int_{|y|_r \geq 1} \mathbb{E} \left\| G(s+t_n, \tilde{X}(s-), y) - \tilde{G}(s, \tilde{X}(s-), y) \right\|^2 v(dy) \right).
 \end{aligned}$$

Using (4.6), it follows that $\zeta_5^n \rightarrow 0$ as $n \rightarrow \infty$, like ζ_1^n . By combining the estimations $J_1 - J_5$, we get for all $t \in \mathbb{R}$

$$\mathbb{E} \left\| X_n(t) - \tilde{X}(t) \right\|^2 \leq \zeta^n + \vartheta \sup_{s \in \mathbb{R}} \mathbb{E} \left\| X_n(s) - \tilde{X}(s) \right\|^2$$

where $\zeta^n = \sum_{i=1}^5 \zeta_i^n$. It follows that

$$\mathbb{E} \left\| X_n(t) - \tilde{X}(t) \right\|^2 \leq \frac{\zeta^n}{1 - \vartheta}.$$

Since $\vartheta < 1$ and $\zeta^n \rightarrow 0$ as $n \rightarrow \infty$ for all $t \in \mathbb{R}$ then one has

$$\mathbb{E} \left\| X_n(t) - \tilde{X}(t) \right\|^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{for each } t \in \mathbb{R}.$$

Hence we deduce that $X(t + t_n)$ converge to $\tilde{X}(t)$ in distribution. Similarly, one can get that $\tilde{X}(t - t_n)$ converge to $X(t)$ in distribution too. Therefore this fixed-point solution of the Eq. (1.1) is square-mean almost automorphic in distribution. which completes the proof. \square

5 Example

Consider the following stochastic integro-differential equations

$$\left\{ \begin{aligned} \frac{\partial u(t,x)}{\partial t} &= \frac{\partial^2}{\partial x^2} u(t,x) + g(t,u(t,x)) + \int_{-\infty}^t e^{-\omega(t-s)} f(s,u(s,x)) ds + \int_{-\infty}^t e^{-\omega(t-s)} h(s,u(s,x)) dW(s) \\ &\quad + \int_{-\infty}^t e^{-\omega(t-s)} \theta(s,u(s,x)) Z(ds), \quad (t,x) \in \mathbb{R} \times (0,1), \\ u(t,0) &= u(t,1) = 0, \quad t \in \mathbb{R} \\ u(0,x) &= u_0(x) = 0, \quad x \in (0,1) \end{aligned} \right. \tag{5.1}$$

where W is a Q -Wiener process on $L^2(0,1)$ with $\text{Tr } Q < \infty$ and Z is a Lévy pure jump process on $L^2(0,1)$ which is independent of W and $\omega > 0$. The forcing terms are follows:

$$g(t,u) = \delta \sin u(\sin t + \sin \sqrt{2}t), \quad f(t,u) = \delta \sin u(\sin t + \sin \sqrt{3}t),$$

$$h(t,u) = \delta \sin u(\sin t + \sin \sqrt{5}t), \quad \theta(t,u) = \delta \sin u(\sin t + \sin \pi t)$$

with $\delta > 0$. Denote $\mathbb{H} = V = L^2(0,1)$. In order to write the system (5.1) on the abstract form (1.1), we consider the linear operator $A : D(A) \subset L^2(0,1) \rightarrow L^2(0,1)$, given by

$$\begin{aligned} D(A) &= H^2(0,1) \cap H_0^1(0,1), \\ Ax(\xi) &= x''(\xi) \quad \text{for } \xi \in (0,1) \quad \text{and } x \in D(A). \end{aligned}$$

It is well-known that A generates a C_0 -semigroup $(T(t))_{t \geq 0}$ on $L^2(0,1)$ defined by

$$(T(t)x)(r) = \sum_{n=1}^{\infty} e^{-n^2\pi^2 t} \langle x, e_n \rangle_{L^2} e_n(r),$$

where $e_n(r) = \sqrt{2} \sin(n\pi r)$ for $n = 1, 2, \dots$, and $\|T(t)\| \leq e^{-\pi^2 t}$ for all $t \geq 0$.

Then the system (5.1) takes the following abstract form

$$\begin{aligned} u'(t) = & Au(t) + g(t, u(t)) + \int_{-\infty}^t B_1(t-s)f(s, u(s))ds \\ & + \int_{-\infty}^t B_2(t-s)h(s, u(s))dW(s) \\ & + \int_{-\infty}^t B_2(t-s) \int_{|y|_V < 1} F(s, u((s-), y)\tilde{N}(ds, dy) \\ & + \int_{-\infty}^t B_2(t-s) \int_{|y|_V \geq 1} G(s, u(s-, y)N(ds, dy) \quad \text{for all } t \in \mathbb{R}, \end{aligned}$$

where $u(t) = u(t, \cdot)$, $B_1(t) = B_2(t) = e^{\omega t}$ for $t \geq 0$ and

$$\theta(t, u)Z(dt) = \int_{|y|_V < 1} F(t, u(t-), y)\tilde{N}(dt, dy) + \int_{|y|_V \geq 1} G(t, u(t-), y)N(dt, dy)$$

with

$$Z(t) = \int_{|y|_V < 1} y\tilde{N}(t, dy) + \int_{|y|_V \geq 1} yN(t, dy),$$

$$F(t, u, y) = h(t, u)y \cdot 1_{\{|y|_V < 1\}} \quad \text{and} \quad G(t, u, y) = h(t, u)y \cdot 1_{\{|y|_V \geq 1\}}.$$

Here, we assume that the Lévy pure jump process Z on $L^2(0, 1)$ is decomposed as above by the Lévy-It decomposition theorem.

Clearly, f, θ are almost automorphic, and F, G Poisson almost automorphic and satisfying H_1-H_3 . Moreover, it is easy to see that the conditions (3.4)–(3.8) are satisfied with

$$\begin{aligned} m_g(t) &= \delta^2 \left(\sin t + \sin \sqrt{2}t \right)^2, & m_f(t) &= \delta^2 \left(\sin t + \sin \sqrt{3}t \right)^2 \\ m_h(t) &= \delta^2 \|Q\|_{L(V,V)} \left(\sin t + \sin \sqrt{5}t \right)^2, & m_F(t) &= \delta^2 \nu(B_1(0)) \left(\sin t + \sin \sqrt{3}t \right)^2 \\ m_G(t) &= \delta^2 b \left(\sin t + \sin \pi t \right)^2, \end{aligned}$$

where $B_1(0)$ is the ball in V centered at the origin with radius 1.

Obviously,

$$\begin{aligned}
L_g &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t e^{-\omega(t-s)} m_g(s) ds = \delta^2 \sup_{t \in \mathbb{R}} \int_{-\infty}^t e^{-\omega(t-s)} m_g(s) ds \left(\sin s + \sin \sqrt{2}s \right)^2 < \infty, \\
L_f &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_1(t-s)\| m_f(s) ds = \delta^2 \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_1(t-s)\| \left(\sin s + \sin \sqrt{3}s \right)^2 ds < \infty, \\
L_h &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\| m_h(s) ds = \delta^2 \|Q\|_{L(V,V)} \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\| \left(\sin s + \sin \sqrt{5}s \right)^2 ds < \infty, \\
L_F &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 m_F(s) ds = \delta^2 v(B_1(0)) \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 \left(\sin s + \sin \pi s \right)^2 ds < \infty, \\
L_G &= \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 m_G(s) ds = \delta^2 b \sup_{t \in \mathbb{R}} \int_{-\infty}^t \|B_2(t-s)\|^2 \left(\sin s + \sin \pi s \right)^2 ds < \infty.
\end{aligned}$$

Therefore, by Theorem 4.1, the Eq. (5.1) has a square-mean almost automorphic in distribution mild solution on \mathbb{R} whenever δ is small enough.

Acknowledgements The author would like to express his sincere gratitude to the referee for his remarks which helped us a lot to improve the original version of this work.

References

1. Applebaum, D. 2009. *Lévy Process and Stochastic Calculus*, 2nd ed. Cambridge: Cambridge University Press.
2. Bezandry, P., and T. Diagana. 2007. Existence of almost periodic solutions to some stochastic differential equations. *Applicable Analysis* 86: 819–827.
3. Bezandry, P. 2008. Existence of almost periodic solutions to some functional integro-differential stochastic evolution equations. *Statistics & Probability Letters* 78: 2844–2849.
4. Bezandry, P. H. and T. Diagana 2007. Square-mean almost periodic solutions nonautonomous stochastic differential equations. *Electronic Journal of Differential Equation* no. 117
5. Bezandry, P., and T. Diagana. 2008. Existence of S^2 -almost periodic solutions to a class of nonautonomous stochastic evolution equations. *Electronic Journal of Qualitative Theory of Differential Equations* 35: 1–19.
6. Bochner, S. 1961. Uniform convergence of monotone sequences of functions. *Proceedings of the National Academy of Sciences of the United States of America* 47: 582–585.
7. Changa, Y.K., Z.H. Zhaoa, G.M. N'Guérékata, and R. Mab. 2011. Stepanov-like almost automorphy for stochastic processes and applications to stochastic differential equations. *Nonlinear Analysis: Real World Applications* 12: 1130–1139.
8. Diagana, T. 2006. Weighted pseudo almost periodic functions and applications. *Comptes Rendus de l'Académie, des Sciences de Paris, Série I* 343 (10): 643–646.
9. Diagana, T. 2008. Weighted pseudo-almost periodic solutions to some differential equations. *Nonlinear Analysis, Theory, Methods and Applications* 68 (8): 2250–2260.
10. Diop, M.A., K. Ezzinbi, and M.M. Mbaye. 2015. Measure theory and S^2 -pseudo almost periodic and automorphic process: Application to stochastic evolution equations. *Afrika Matematika* 26 (5): 779–812.
11. Diop, M.A., K. Ezzinbi, and M.M. Mbaye. 2015. Existence and global attractiveness of a pseudo almost periodic solution in the p -th mean sense for stochastic evolution equation driven by a fractional Brownian. *Stochastics: An International Journal of Probability and Stochastic Processes* 87 (6): 1061–1093.
12. Diop, M.A., K. Ezzinbi, and M.M. Mbaye. 2017. Existence and global attractiveness of a square-mean μ -pseudo almost automorphic solution for some stochastic evolution equation driven by Lévy noise. *Mathematische Nachrichten* 290 (8–9): 1260–1280.
13. Fu, M.M. 2012. Almost automorphic solutions for nonautonomous stochastic differential equations. *Journal of Mathematical Analysis and Applications* 393: 231–238.
14. Fu, M.M., and Z.X. Liu. 2010. Square-mean almost automorphic solutions for some stochastic differential equations. *Proceedings of the American Mathematical Society* 133: 3689–3701.

15. Kamenskii, M., O. Mellah, and P. Raynaud de Fitte. 2015. Weak averaging of semilinear stochastic differential equations with almost periodic coefficients. *Journal of Mathematical Analysis and Applications* 427: 336–364.
16. Bedouhene, F., N. Challali, O. Mellah, P.R. Fitte, and M. Smaali. 2015. Almost automorphy and various extensions for stochastic processes. *Journal of Mathematical Analysis and Applications* 429 (2): 1113–1152.
17. Mellah, O., and P. Raynaud de Fitte. 2013. Counterexamples to mean square almost periodicity of the solutions of some SDEs with almost periodic coefficients. *Electronic Journal of Differential Equations* 2013 (91): 1–7.
18. Liu, Z., and K. Sun. 2014. Almost automorphic solutions for stochastic differential equations driven by Lévy noise. *Journal of Functional Analysis* 266 (3): 1115–1149.
19. Mbaye, M.M. 2017. Square-mean μ -pseudo almost periodic and automorphic solutions for a class of semilinear integro-differential stochastic evolution equations. *Afrika Matematika* 28 (3–4): 643–660.
20. Mbaye, M.M. 2016. Almost automorphic solution for some stochastic evolution equation driven by Lévy noise with coefficients S^2 -almost automorphic. *Nonautonomous Dynamical Systems* 3: 85–103.
21. Morozan, T., and C. Tudor. 1989. Almost periodic solutions of affine itô equations. *Stochastic Analysis and Applications* 7 (4): 451–474.
22. N'Guérékata, G.M. 2009. Almost automorphic solutions to second-order semilinear evolution equations. *Nonlinear Analysis, Theory, Methods and Applications* 71 (12): 432–435.
23. N'Guérékata, G.M. 2005. *Topics in almost automorphy*. New York: Springer.
24. Peszat, S., and J. Zabczyk. 2007. *Stochastic Partial Differential Equations with Lévy Noise*. Cambridge: Cambridge University Press.
25. Prato, G.D., and J. Zabczyk. 1992. *Stochastic Equations in Infinite Dimensions, Encyclopedia of Mathematics and its Applications* 44. Cambridge: Cambridge University Press.
26. Tudor, C. 1992. Almost periodic solutions of affine stochastic evolutions equations. *Stochastics and Stochastics Reports* 38: 251–266.
27. Xia, Z. Pseudo almost automorphic in distribution solutions of semilinear stochastic integro-differential equations by measure theory. *International Journal of Mathematics* <https://doi.org/10.1142/S0129167X15501128>
28. Wang, Y., and Z.X. Liu. 2012. Almost periodic solutions for stochastic differential equations with Lévy noise. *Nonlinearity* 25: 2803–2821.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Asymptotically Periodic Solution of a Stochastic Differential Equation

Solym Mawaki Manou-Abi^{1,2} · William Dimbour³

Received: 23 April 2018 / Revised: 10 December 2018

© Malaysian Mathematical Sciences Society and Penerbit Universiti Sains Malaysia 2019

Abstract

In this paper, we first introduce the concept and properties of ω -periodic limit process. Then, we apply specific criteria obtained to investigate asymptotically ω -periodic mild solutions of a Stochastic differential equation driven by a Brownian motion. Finally, we give an example to show usefulness of the theoretical results that we obtained in the paper.

Keywords Square mean asymptotically periodic · Square mean periodic limit · Stochastic differential equation · Semigroup mild solution

Mathematics Subject Classification 34C25 · 34C27 · 60H30 · 34 F05

1 Introduction

The recurrence of dynamics for stochastic and deterministic processes produced by many different kinds of stochastic and deterministic equations is one of the most important topics in the qualitative theory of stochastic processes and functions, due to both its mathematical interest and its applications in many scientific fields, such as mathematical biology, celestial mechanics, non linear vibration, control theory. The

Communicated by Shangjiang Guo.

Solym Mawaki Manou-Abi
solym.manou-abi@univ-mayotte.fr; solym-mawaki.manou-abi@umontpellier.fr

William Dimbour
william.dimbour@espe-guyane.fr

¹ Département Sciences et Technologies, CUFR de Mayotte, Dembeni, France

² Institut Montpelliérain Alexander Grothendieck, UMR CNRS 5149, Université de Montpellier, Montpellier, France

³ UMR Espace-Dev, Université de Guyane, Campus de Troubiran, 97300 Cayenne, Guyane (FWI), France

concept of periodicity was studied for dynamics of stochastic processes and functions. However, the dynamics observed in some phenomena in the real world are not periodic, but almost approximately or asymptotically periodic; see, for instance, [1–3] for almost periodic observations.

In the past several decades, many authors suggested and developed several extensions of the concept of periodicity, in the deterministic and stochastic case, such as the almost automorphy, pseudo almost periodicity, asymptotically periodicity. (see [4–17] and references therein)

Recently, the concept of periodic limit function has been introduced by Xie and Zhang [18] to generalize the notion of asymptotic periodicity. The authors investigate some properties of periodic limit functions in order to study the existence and uniqueness of asymptotically periodic solutions of some differential equations. However, to the best of our knowledge, there is no work or applicable results for stochastic differential equations. Therefore, in this paper, we will introduce the notion of square mean periodic limit process. Then we will investigate their qualitative properties in order to study the existence of square mean asymptotically periodic mild solution to the following stochastic differential equation (SDE) driven by Brownian motion :

$$\begin{cases} dX(t) = AX(t)dt + f(t, X(t))dt + g(t, X(t))dB(t), & t \geq 0 \\ X(0) = c_0, \end{cases}$$

where $c_0 \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ and A is an infinitesimal generator which generates a C_0 semi-group, denoted by $(T(t))_{t \geq 0}$. In addition,

$$\begin{aligned} f &: \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H}), \\ g &: \mathbb{R} \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \end{aligned}$$

are Lipschitz continuous and bounded, and $B(t)$ is a two-sided standard one-dimensional Brownian motion, which is defined on the filtered complete probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ with values in the separable Hilbert space \mathbb{H} . Here, $\mathcal{F}_t = \sigma\{B(u) - B(v)/u, v \leq t\}$.

The paper is organized as follows: In Sect. 2, we preliminarily introduce the space of square mean ω -periodic limit process and study properties of such processes. It also includes some results, not only on the completeness of the space that consists of the square mean ω -periodic limit processes, but also on the composition of such processes. Based on the results in Sect. 2 and given some suitable conditions, we prove in Sect. 3 the existence as well as the uniqueness of the square mean asymptotically ω -periodic solution for the above SDE. Finally, an illustrative example is provided to show the feasibility of the theoretical results developed in the paper.

2 Square Mean Omega-Periodic Limit Process

This section is concerned with some notations, definitions, lemmas and preliminary facts that may be used in what follows. Throughout this paper, we consider a real separable Hilbert space $(\mathbb{H}, \|\cdot\|)$ and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with

a filtration $(\mathcal{F}_t)_t$. Denote by $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ the space of all strongly measurable square integrable \mathbb{H} -valued random variables such that

$$\mathbb{E}\|X\|^2 = \int_{\Omega} \|X(\omega)\|^2 d\mathbb{P}(\omega) < \infty.$$

For $X \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$, let $\|X\|_2 = (\mathbb{E}\|X\|^2)^{1/2}$. Then, it is routine to check that $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is a Hilbert space equipped with the norm $\|\cdot\|_2$.

Definition 2.1 A stochastic process $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is said to be continuous whenever

$$\lim_{t \rightarrow s} \mathbb{E}\|X(t) - X(s)\|^2 = 0.$$

Definition 2.2 A stochastic process $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is said to be bounded if there exists a constant $K > 0$ such that

$$\mathbb{E}\|X(t)\|^2 \leq K \quad \forall t \geq 0$$

By $CUB(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, we denote the collection of all continuous and uniformly bounded stochastic processes from $\mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Definition 2.3 A continuous and bounded stochastic process $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is said to be square mean ω -periodic limit if there exists $\omega > 0$ such that

$$\lim_{n \rightarrow +\infty} \mathbb{E}\|X(t + n\omega) - \tilde{X}(t)\|^2 = 0$$

is well defined for each $t \geq 0$ when $n \in \mathbb{N}$ for some stochastic process $\tilde{X} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Remark 2.1 For all $t \geq 0$, $\tilde{X}(t)$ is the limit of $X(t + n\omega)$ in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ when $n \rightarrow +\infty$, when it exists. The collection of such ω -periodic limit processes is denoted by $P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. Note also that the process \tilde{X} in the previous definition is measurable but not necessarily continuous.

In the following Proposition, we list some properties of square mean ω -periodic limit process.

Proposition 2.1 *Let X be square mean ω -periodic limit process such that*

$$\lim_{n \rightarrow +\infty} \mathbb{E}\|X(t + n\omega) - \tilde{X}(t)\|^2 = 0$$

is well defined for each $t \geq 0$ when $n \in \mathbb{N}$ for some stochastic process $\tilde{X} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

If X, X_1, X_2 are square mean ω -periodic limit processes, then the following are true :

- (a) $X_1 + X_2$ is a square mean ω -periodic limit process.
- (b) cX is a square mean ω -periodic limit process for every scalar c .
- (c) We have

$$\mathbb{E} \|\tilde{X}(t + \omega) - \tilde{X}(t)\|^2 = 0.$$

- (d) \tilde{X} is bounded on \mathbb{R}_+ and $\|\tilde{X}\|_\infty \leq \|X\|_\infty \leq K$.
- (e) $X_a(t) = X(t + a)$ is a square mean ω -periodic limit process for each fixed $a \in \mathbb{R}_+$.

Proof The proof is straightforward, but we will only prove the statement in (c). To this end, note that for each $n \geq 1$, we have

$$0 \leq \mathbb{E} \left\| \tilde{X}(t + \omega) - \tilde{X}(t) \right\|^2 \leq 2\mathbb{E} \left\| \tilde{X}(t + \omega) - X(t + (n + 1)\omega) \right\|^2 + 2\mathbb{E} \left\| X(t + (n + 1)\omega) - \tilde{X}(t) \right\|^2,$$

Let $\epsilon > 0$, for N sufficiently large, if $n \geq N$, then

$$\mathbb{E} \left\| \tilde{X}(t + \omega) - X(t + (n + 1)\omega) \right\|^2 \leq \epsilon/2$$

and

$$\mathbb{E} \left\| X(t + (n + 1)\omega) - \tilde{X}(t) \right\|^2 \leq \epsilon/2.$$

so that $\mathbb{E} \left\| \tilde{X}(t + \omega) - \tilde{X}(t) \right\|^2 \leq \epsilon$. Thus $\mathbb{E} \left\| \tilde{X}(t + \omega) - \tilde{X}(t) \right\|^2 = 0$. □

Because of the above Proposition, we give name of ω -periodic limit process in Definition 2.3.

Remark 2.2 Note that if

$$\mathbb{E} \|\tilde{X}(t + \omega) - \tilde{X}(t)\|^2 = 0$$

then

$$\mathbb{E} \|\tilde{X}(t + p\omega) - \tilde{X}(t)\|^2 = 0 \text{ for all } p \geq 1.$$

Theorem 2.1 The space $P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is a Banach space equipped with the norm $\|X\|_\infty = \sup_{t \geq 0} (\mathbb{E} \|X(t)\|^2)^{1/2} = \sup_{t \geq 0} \|X(t)\|_2$.

Proof By Proposition 2.1, $P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is a vector space, and then, it is easy to verify that $\|\cdot\|_\infty$ is a norm on $P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. We only need to show that $P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is complete with respect to the norm $\|\cdot\|_\infty$. To this end, assume

that $(X_n)_{n \geq 0} \in P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is a Cauchy sequence with respect to $\|\cdot\|_\infty$ and that X is the pointwise limit of X_n with respect to $\|\cdot\|_2$; i.e.,

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|X_n(t) - X(t)\|^2 = 0 \tag{1}$$

for each $t \geq 0$. Note that this limit X always exists by the completeness of $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ with respect to $\|\cdot\|_2$.

Since $(X_n)_{n \geq 0}$ is Cauchy with respect to $\|\cdot\|_\infty$, the convergence in (1) is uniform for $t \geq 0$. We need to show that $X \in P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. First note that X is stochastically continuous from the uniform convergence of X_n to X with respect to $\|\cdot\|_2$ and the stochastic continuity of X_n . Next, we prove that X is a square mean ω -periodic limit process. By the definition of $(X_n)_{n \geq 0} \in P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, we have for all $i \geq 0$, :

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|X_i(t + n\omega) - \tilde{X}_i(t)\|^2 = 0, \tag{2}$$

for some stochastic process $\tilde{X}_i : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$. Let us point out that for each $t \geq 0$, the sequence $(\tilde{X}_i(t))_{i \geq 0}$ is a Cauchy sequence in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$. Indeed, we have

$$\begin{aligned} \mathbb{E} \left\| \tilde{X}_i(t) - \tilde{X}_k(t) \right\|^2 &\leq 3 \mathbb{E} \left\| \tilde{X}_i(t) - X_i(t + n\omega) \right\|^2 \\ &\quad + 3 \mathbb{E} \|X_i(t + n\omega) - X_k(t + n\omega)\|^2 \\ &\quad + 3 \mathbb{E} \left\| X_k(t + n\omega) - \tilde{X}_k(t) \right\|^2. \end{aligned}$$

By (2) and the fact that the sequence $(X_n)_{n \geq 0}$ is Cauchy, we get that the sequence $(\tilde{X}_i(t))_{i \geq 0}$ is Cauchy.

Using the completeness of the space $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$, we denote by \tilde{X} the pointwise limit of $(\tilde{X}_i(t))_{i \geq 0}$ such that

$$\lim_{i \rightarrow +\infty} \mathbb{E} \|\tilde{X}_i(t) - \tilde{X}(t)\|^2 = 0. \tag{3}$$

Let us prove now that

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|X(t + n\omega) - \tilde{X}(t)\|^2 = 0,$$

for each $t \geq 0$. Indeed for each $i \geq 0$, we have

$$\begin{aligned} \mathbb{E} \left\| X(t + n\omega) - \tilde{X}(t) \right\|^2 &\leq 3\mathbb{E} \|X(t + n\omega) - X_i(t + n\omega)\|^2 \\ &\quad + 3\mathbb{E} \left\| X_i(t + n\omega) - \tilde{X}_i(t) \right\|^2 \\ &\quad + 3\mathbb{E} \left\| \tilde{X}_i(t) - \tilde{X}(t) \right\|^2. \end{aligned}$$

By (1), (2) and (3), we get

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|X(t + n\omega) - \tilde{X}(t)\|^2 = 0.$$

The proof is completed. □

Definition 2.4 A continuous and bounded process $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is said to be square mean asymptotically ω -periodic if $X = Y + Z$ where Y and Z are continuous bounded processes such that

$$\mathbb{E} \|Y(t + \omega) - Y(t)\|^2 = 0 \quad \text{and} \quad \lim_{t \rightarrow \infty} \mathbb{E} \|Z(t)\|^2 = 0.$$

We write $Y \in P_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, $Z \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, and we denote the space of all square mean asymptotically ω -periodic stochastic process $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ by $AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$.

Lemma 2.2 *The space $AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is a closed subspace of $P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$*

Proof Note that

$$AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H})) \subseteq P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H})).$$

Indeed, if $X \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, then we have for all $n \geq 1, t \geq 0$,

$$\begin{aligned} \mathbb{E} \|X(t + n\omega) - Y(t)\|^2 &\leq 2\mathbb{E} \|X(t + n\omega) - Y(t + n\omega)\|^2 + 2\mathbb{E} \|Y(t + n\omega) - Y(t)\|^2 \\ &\leq 2\mathbb{E} \|Z(t + n\omega)\|^2 + 2\mathbb{E} \|Y(t + n\omega) - Y(t)\|^2 \\ &= 2\mathbb{E} \|Z(t + n\omega)\|^2 \end{aligned}$$

so that

$$\lim_{n \rightarrow \infty} \mathbb{E} \|X(t + n\omega) - Y(t)\|^2 = 0$$

for all $t \geq 0$.

Now let us show that $AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is a closed space.

Let $X \in \overline{AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))}$; there exists $X_n \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ such that $\lim_{n \rightarrow \infty} X_n = X$. Since $X_n \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, we have $X_n = Y_n + Z_n$ where $Y_n \in P_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ and $Z_n \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$.

If Y_n or Z_n does not converge, then X_n will not converge. Thus, there exist Y and Z such that $\lim_{n \rightarrow \infty} Y_n = Y$ and $\lim_{n \rightarrow \infty} Z_n = Z$. We have $X = Y + Z$.

In the sequel, we will show that $Y \in P_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ and $Z \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. Firstly, we have

$$\begin{aligned} & \mathbb{E} \|Y(t + \omega) - Y(t)\|^2 \\ & \leq 3\mathbb{E} \|Y(t + \omega) - Y_n(t + \omega)\|^2 + 3\mathbb{E} \|Y_n(t + \omega) - Y_n(t)\|^2 \\ & \quad + 3\mathbb{E} \|Y_n(t) - Y(t)\|^2 \\ & = 3\mathbb{E} \|Y(t + \omega) - Y_n(t + \omega)\|^2 + 3\mathbb{E} \|Y_n(t) - Y(t)\|^2 \end{aligned}$$

For N sufficiently large, if $n \geq N$, then

$$\begin{aligned} & \mathbb{E} \|Y(t + \omega) - Y_n(t + \omega)\|^2 \leq \epsilon/6 \\ & \mathbb{E} \|Y_n(t) - Y(t)\|^2 \leq \epsilon/6. \end{aligned}$$

Therefore, for $n > N$,

$$E \|Y(t + \omega) - Y(t)\|^2 \leq \epsilon.$$

Thus

$$E \|Y(t + \omega) - Y(t)\|^2 = 0,$$

so that $Y \in P_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$.

On the other hand,

$$\mathbb{E} \|Z(t)\|^2 \leq 2\mathbb{E} \|Z(t) - Z_n(t)\|^2 + 2\mathbb{E} \|Z_n(t)\|^2.$$

For all $\epsilon > 0, \exists T_\epsilon > 0, t > T_\epsilon \Rightarrow$

$$\mathbb{E} \|Z_n(t)\|^2 \leq \epsilon/4.$$

There exists $N \in \mathbb{N}, n > N \Rightarrow$

$$E \|Z(t) - Z_n(t)\|^2 \leq \epsilon/4.$$

Therefore, for all $n > N$ and $t > T_\epsilon$, we have

$$\mathbb{E} \|Z(t)\|^2 \leq \epsilon/2 + \epsilon/2 = \epsilon,$$

so that $Z \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. □

From the above Lemma, we have the following conclusion by the fundamental knowledge of functional analysis.

Theorem 2.3 *The space $AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ is a Banach space equipped with the norm $\|\cdot\|_\infty$.*

The following result provides some interesting properties.

Theorem 2.4 *Let X be a continuous and bounded stochastic process and $\omega > 0$. Then, the following statements are equivalent*

- (i) $X \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$
- (ii) We have

$$\lim_{n \rightarrow \infty} \mathbb{E} \|X(t + n\omega) - Y(t)\|^2 = 0$$

uniformly on $t \in \mathbb{R}_+$ for some stochastic process $Y : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

- (iii) We have

$$\lim_{n \rightarrow \infty} \mathbb{E} \|X(t + n\omega) - Y(t)\|^2 = 0$$

uniformly on compact subsets of \mathbb{R}_+ for some stochastic process $Y : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

- (iv) We also have

$$\lim_{n \rightarrow \infty} \mathbb{E} \|X(t + n\omega) - Y(t)\|^2 = 0$$

uniformly on $[0, \omega]$ for some stochastic process $Y : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Proof Clearly, Statement (ii) implies (iii) and (iii) implies (iv). Suppose that (i) holds and let $X = Y + Z$ where $Y \in P_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ and $Z \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. Now for $n \in \mathbb{N}$,

$$X(t + n\omega) = Y(t + n\omega) + Z(t + n\omega) \quad (\star). \tag{4}$$

Let $\epsilon > 0$. Since $Z \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, there exists N_1 such that $\mathbb{E}\|Z(t + n\omega)\|^2 < \epsilon/2$ whenever $n \geq N_1$ for every $t \in \mathbb{R}_+$. Then, using (4), we obtain

$$\begin{aligned} \mathbb{E}\|X(t + n\omega) - Y(t)\|^2 &\leq 2\mathbb{E}\|X(t + n\omega) - Y(t + n\omega)\|^2 \\ &\quad + 2\mathbb{E}\|Y(t + n\omega) - Y(t)\|^2 \\ &= 2\mathbb{E}\|Z(t + n\omega)\|^2 \\ &\leq \epsilon, \end{aligned}$$

whenever $n \geq N_1$ for every $t \in \mathbb{R}_+$. This shows that

$$\lim_{n \rightarrow \infty} \mathbb{E} \|X(t + n\omega) - Y(t)\|^2 = 0$$

uniformly on $t \geq 0$ for some stochastic process $Y : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Hence, (i) implies (ii).

Finally, suppose that (iv) holds. It is clear that Y is bounded on \mathbb{R}_+ and

$$\mathbb{E} \|Y(t + \omega) - Y(t)\|^2 = 0$$

for each $t \geq 0$ like in Proposition 2.1, part (c).

Thus, to show the continuity of Y on \mathbb{R}_+ , we only need to prove that Y is continuous on $[0, \omega]$. Now, take any fixed $t_0 \in [0, \omega]$ and let $t \in [0, \omega]$. For each $n \in \mathbb{N}$, we have

$$\mathbb{E} \|Y(t) - Y(t_0)\|^2 \leq 3\mathbb{E} \|Y(t) - X(t + n\omega)\|^2 \tag{5}$$

$$+ 3\mathbb{E} \|X(t + n\omega) - X(t_0 + n\omega)\|^2 \tag{6}$$

$$+ 3\mathbb{E} \|X(t_0 + n\omega) - Y(t_0)\|^2 \tag{7}$$

Let $\epsilon > 0$. By Assumption in (iv), we conclude that there exists a positive integer N_2 such that

$$\mathbb{E} \|Y(t) - X(t + n\omega)\|^2 < \epsilon/9 \tag{8}$$

for $t \in [0, \omega]$ and $n \geq N_2$.

On the other hand, since $X(t + N_2\omega)$ is in $C_b(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, then there exists $\delta > 0$ such that

$$\mathbb{E} \|X(t + N_2\omega) - X(t_0 + N_2\omega)\|^2 < \epsilon/9 \tag{9}$$

for $|t - t_0| < \delta$.

Using (5)–(9), we conclude that

$$\mathbb{E} \|Y(t) - Y(t_0)\|^2 < \epsilon \text{ when } |t - t_0| < \delta,$$

which show that Y is continuous on $[0, \omega]$. Hence, $Y \in P_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$.

Next, we will show that $X - Y \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. Suppose that $\epsilon > 0$ and there exists a positive integer N_3 such that

$$\mathbb{E} \|X(t + n\omega) - Y(t)\|^2 < \epsilon$$

when $n \geq N_3$ uniformly for $t \in [0, \omega]$ by Assumption in (iv) again.

Thus, for $n = N_3 + k, k = 0, 1, 2, \dots$, we conclude that

$$\mathbb{E} \|X(t + (N_3 + k)\omega) - Y(t)\|^2 < \epsilon$$

uniformly for $t \in [0, \omega]$. Moreover, if we denote $t' = t + k\omega$, where $t' \in [k\omega, (k+1)\omega]$, $t \in [0, \omega]$ and $k = 0, 1, 2, \dots$, then we obtain

$$\begin{aligned} & \mathbb{E} \left\| X(t' + N_3\omega) - Y(t' + N_3\omega) \right\|^2 \\ &= \mathbb{E} \left\| X(t + (N_3 + k)\omega) - Y(t + (N_3 + k)\omega) \right\|^2 \\ &= \mathbb{E} \left\| X(t + (N_3 + k)\omega) - Y(t) \right\|^2 < \epsilon \end{aligned}$$

for $t' \in [k\omega, (k + 1)\omega], k = 0, 1, 2, \dots$

That is

$$\mathbb{E} \left\| X(t) - Y(t) \right\|^2 < \epsilon \quad (t \geq N_3\omega)$$

which show that $X - Y \in C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$. Hence, $X \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ and (iv) implies (i). \square

The following generalizes the Definition 2.3.

Definition 2.5 A continuous bounded process $f : \mathbb{R}_+ \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is called square mean ω periodic limit in $t \in \mathbb{R}_+$ uniformly for X in bounded sets of $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ if for every bounded subsets K of $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$, $\{f(t, X) : t \in \mathbb{R}_+; X \in K\}$ is bounded and

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left\| f(t + n\omega, X) - \tilde{f}(t, X) \right\|^2 = 0$$

is well defined when $n \in \mathbb{N}$ for each $t \geq 0$ and for some process $\tilde{f} : \mathbb{R}_+ \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

We have the following composition result:

Theorem 2.5 Assume that $f : \mathbb{R}_+ \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ is a square mean ω -periodic limit process uniformly for $Y \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ in bounded sets of $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ and satisfies the Lipschitz condition, that is, there exists constant $L > 0$ such that

$$\mathbb{E} \left\| f(t, Y) - f(t, Z) \right\|^2 \leq L \mathbb{E} \left\| Y - Z \right\|^2 \quad \forall t \geq 0, \forall Y, Z \in \mathbb{L}^2(\mathbb{P}, \mathbb{H}).$$

Let $X : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ be a square mean ω -periodic limit process. Then, the process $F(t) = (f(t, X(t)))_{t \geq 0}$ is a square mean ω -periodic limit process.

Proof Since X is a square mean ω -periodic limit process, we have

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left\| X(t + n\omega) - \tilde{X}(t) \right\|^2 = 0 \tag{10}$$

for some stochastic process $\tilde{X} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

By using Proposition 2.1 (4), we can choose a bounded subset K of $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ such that $X(t), \tilde{X}(t) \in K$ for all $t \geq 0$. Then, $F(t)$ is bounded.

On the other hand, we have

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left\| f(t + n\omega, X) - \tilde{f}(t, X) \right\|^2 = 0, \tag{11}$$

for each $t \geq 0$ and each $X \in K$.

Let us consider the process $\tilde{F} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ defined by $\tilde{F}(t) = \tilde{f}(t, \tilde{X}(t))$. Note that

$$\begin{aligned} \mathbb{E} \left\| F(t + n\omega) - \tilde{F}(t) \right\|^2 &= \mathbb{E} \left\| f(t + n\omega, \tilde{X}(t + n\omega)) - \tilde{f}(t, \tilde{X}(t)) \right\|^2 \\ &\leq 2\mathbb{E} \left\| f(t + n\omega, X(t + n\omega)) - f(t + n\omega, \tilde{X}(t)) \right\|^2 \\ &\quad + 2\mathbb{E} \left\| f(t + n\omega, \tilde{X}(t)) - \tilde{f}(t, \tilde{X}(t)) \right\|^2 \\ &\leq 2L \mathbb{E} \left\| X(t + n\omega) - \tilde{X}(t) \right\|^2 \\ &\quad + 2\mathbb{E} \left\| f(t + n\omega, \tilde{X}(t)) - \tilde{f}(t, \tilde{X}(t)) \right\|^2 \end{aligned}$$

We deduce from (10) and (11) that

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left\| F(t + n\omega) - \tilde{F}(t) \right\|^2 = 0$$

is well defined for $\tilde{F}(t) = \tilde{f}(t, \tilde{X}(t))$. □

Now, let us end this section with the following property of a Brownian motion.

Proposition 2.2 (Weak Markov Property). *Let $B = (B(s))_{s \geq 0}$ be a two-sided standard one-dimensional Brownian motion and set for $h \in \mathbb{R}$,*

$$\tilde{B}^h(u) = B(u + h) - B(h), u \in \mathbb{R}.$$

Then, the process \tilde{B}^h is a two-sided Brownian motion independent of $\{B(s) : s \leq h\}$. In others words, $B(u + h)$ has the same law as $\tilde{B}^h(u) + B(h)$.

3 A Stochastic Differential Equation

In this section, we investigate the existence of the square mean asymptotically ω -periodic solution to the following SDE :

$$\begin{cases} dX(t) = AX(t)dt + f(t, X(t))dt + g(t, X(t))dB(t), & t \geq 0 \\ X(0) = c_0, \end{cases} \quad (12)$$

where A is a closed linear operator and

$$\begin{aligned} f &: \mathbb{R}_+ \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H}), \\ g &: \mathbb{R}_+ \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \end{aligned}$$

are Lipschitz continuous and bounded, $(B(t))_t$ is a two-sided standard one-dimensional Brownian motion with values in \mathbb{H} and \mathcal{F}_t -adapted and $c_0 \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$. Recall that $\mathcal{F}_t = \sigma\{B(u) - B(v)/u, v \leq t\}$.

In order to establish our main result, we impose the following conditions.

(H1): A generates an exponentially stable semigroup $(T(t))_{t \geq 0}$ in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$, that is, a linear operator, such that :

1. $T(0) = I$ where I is the identity operator.
2. $T(t)T(r) = T(t+r)$ for all $t, r \geq 0$.
3. The map $t \mapsto T(t)x$ is continuous for every fixed $x \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.
4. There exist $M > 0$ and $a > 0$ such that $\|T(t)\| \leq Me^{-at}$ for $t \geq 0$.

Definition 3.1 The \mathcal{F}_t -progressively measurable process $\{X(t), t \geq 0\}$ is said to be a mild solution of (12) if it satisfies the following stochastic integral equation:

$$X(t) = T(t)c_0 + \int_0^t T(t-s)f(s, X(s))ds + \int_0^t T(t-s)g(s, X(s))dB(s).$$

Now, we will establish some technical results.

Lemma 3.1 Let F be a square mean ω -periodic limit process in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$. Under Assumption (H1), the sequence of stochastic processes $(X_n(t))_{n \geq 1}, t \geq 0$, defined by

$$X_n(t) = \int_0^{n\omega} T(t+s)F(n\omega-s)ds$$

is a Cauchy sequence in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ for all $t \geq 0$.

We shall denote by $U = (U(t))_{t \geq 0}$ the limit process of $(X_n(t))_{n \geq 1}, t \geq 0$, in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Proof We have

$$\begin{aligned} & \mathbb{E}\|X_{n+p}(t) - X_n(t)\|^2 \\ &= \mathbb{E}\left\| \int_0^{(n+p)\omega} T(t+s)F((n+p)\omega-s)ds - \int_0^{n\omega} T(t+s)F(n\omega-s)ds \right\|^2 \\ &\leq 2\mathbb{E}\left\| \int_{n\omega}^{(n+p)\omega} T(t+s)F((n+p)\omega-s)ds \right\|^2 \\ &\quad + 2\mathbb{E}\left\| \int_0^{n\omega} T(t+s)(F((n+p)\omega-s) - F(n\omega-s))ds \right\|^2 \\ &= I_1(t, n, p) + I_2(t, n, p) \end{aligned}$$

where

$$I_1(t, n, p) = 2\mathbb{E} \left\| \int_{n\omega}^{(n+p)\omega} T(t+s)F((n+p)\omega - s)ds \right\|^2$$

$$I_2(t, n, p) = 2\mathbb{E} \left\| \int_0^{n\omega} T(t+s) (F((n+p)\omega - s) - F(n\omega - s)) ds \right\|^2.$$

We have

$$\begin{aligned} I_1(t, n, p) &= 2\mathbb{E} \left\| \int_{n\omega}^{(n+p)\omega} T(t+s)F((n+p)\omega - s)ds \right\|^2 \\ &\leq 2\mathbb{E} \left(\int_{n\omega}^{(n+p)\omega} \|T(t+s)F((n+p)\omega - s)\| ds \right)^2 \\ &\leq 2\mathbb{E} \left(\int_{n\omega}^{(n+p)\omega} \|T(t+s)\| \|F((n+p)\omega - s)\| ds \right)^2 \\ &\leq 2\mathbb{E} \left(\int_{n\omega}^{(n+p)\omega} M e^{-a(t+s)} \|F((n+p)\omega - s)\| ds \right)^2 \\ &\leq 2 \int_{n\omega}^{(n+p)\omega} M^2 e^{-2a(t+s)} ds \int_{n\omega}^{(n+p)\omega} \mathbb{E} \|F((n+p)\omega - s)\|^2 ds \\ &\leq 2M^2 p\omega K \int_{n\omega}^{+\infty} e^{-2as} ds \\ &\leq \frac{2M^2 p\omega K}{2a} e^{-2an\omega}. \end{aligned}$$

Now, we consider the integers N_1 and N_2 such that

$$\frac{M^2 p\omega K}{a} e^{-2an\omega} \leq \epsilon \quad \forall n \geq N_1$$

$$\frac{8M^2 K}{a^2} e^{-2an\omega} < \epsilon \quad \forall n \geq N_2$$

and set $N = \max(N_1, N_2)$. For $n \geq N$, we have :

$$\begin{aligned} I_2(t, n, p) &= 2\mathbb{E} \left\| \int_0^{n\omega} T(t+s) (F((n+p)\omega - s) - F(n\omega - s)) ds \right\|^2 \\ &\leq 2\mathbb{E} \left(\int_0^{n\omega} \|T(t+s)\| \|F((n+p)\omega - s) - F(n\omega - s)\| ds \right)^2 \\ &\leq 4\mathbb{E} \left(\int_0^{N\omega} \|T(t+s)\| \|F((n+p)\omega - s) - F(n\omega - s)\| ds \right)^2 \end{aligned}$$

$$\begin{aligned}
 & + 4\mathbb{E} \left(\int_{N\omega}^{n\omega} \|T(t+s)\| \|F((n+p)\omega-s) - F(n\omega-s)\| ds \right)^2 \\
 & \leq I_3(t, n, p) + I_4(t, n, p)
 \end{aligned}$$

where

$$\begin{aligned}
 I_3(t, n, p) & = 4\mathbb{E} \left(\int_0^{N\omega} \|T(t+s)\| \|F((n+p)\omega-s) - F(n\omega-s)\| ds \right)^2 \\
 I_4(t, n, p) & = 4\mathbb{E} \left(\int_{N\omega}^{n\omega} \|T(t+s)\| \|F((n+p)\omega-s) - F(n\omega-s)\| ds \right)^2.
 \end{aligned}$$

From $F \in P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, there exists a stochastic process $\tilde{F} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ such that

$$\lim_{k \rightarrow +\infty} \mathbb{E} \|F(t+k\omega) - \tilde{F}(t)\|^2 = 0$$

is well defined for each $t \geq 0$ when $k \in \mathbb{N}$.

Now, we have :

$$\begin{aligned}
 & I_3(t, n, p) \\
 & \leq 8\mathbb{E} \left(\int_0^{N\omega} M e^{-a(t-s+N\omega)} \|F((n-N+p)\omega+s) - \tilde{F}(s)\| ds \right)^2 \\
 & \quad + 8\mathbb{E} \left(\int_0^{N\omega} M e^{-a(t-s+N\omega)} \|F((n-N)\omega+s) - \tilde{F}(s)\| ds \right)^2,
 \end{aligned}$$

and by Cauchy Schwarz inequality, we obtain

$$\begin{aligned}
 & I_3(t, n, p) \\
 & \leq 8 \int_0^{N\omega} M^2 e^{-2a(t-s+N\omega)} ds \int_0^{N\omega} \mathbb{E} \|F((n-N+p)\omega+s) - \tilde{F}(s)\|^2 ds \\
 & \quad + 8 \int_0^{N\omega} M^2 e^{-2a(t-s+N\omega)} ds \int_0^{N\omega} \mathbb{E} \|F((n-N)\omega+s) - \tilde{F}(s)\|^2 ds \\
 & \leq 8M^2 \int_0^{N\omega} e^{-2a(-s+N\omega)} ds \int_0^{N\omega} \mathbb{E} \|F((n-N+p)\omega+s) - \tilde{F}(s)\|^2 ds \\
 & \quad + 8M^2 \int_0^{N\omega} e^{-2a(-s+N\omega)} ds \int_0^{N\omega} \mathbb{E} \|F((n-N)\omega+s) - \tilde{F}(s)\|^2 ds \\
 & \leq \frac{8M^2}{2a} \int_0^{N\omega} \mathbb{E} \|F((n-N+p)\omega+s) - \tilde{F}(s)\|^2 ds \\
 & \quad + \frac{8M^2}{2a} \int_0^{N\omega} \mathbb{E} \|F((n-N)\omega+s) - \tilde{F}(s)\|^2 ds
 \end{aligned}$$

Using the fact that

$$\max\{\mathbb{E}\|F((n - N + p)\omega + s) - \tilde{F}(s)\|^2, \mathbb{E}\|F((n - N)\omega + s) - \tilde{F}(s)\|^2\} \leq 2K,$$

it follows that

$$\lim_{n \rightarrow \infty} I_3(t, n, p) = 0 \quad \text{for all } t \geq 0,$$

by Lebesgue's dominated convergence theorem.

Similarly,

$$\begin{aligned} I_4(t, n, p) &= 4\mathbb{E} \left\| \int_{N\omega}^{n\omega} T(t+s) (F((n+p)\omega - s) - F(n\omega - s)) ds \right\|^2 \\ &\leq 4\mathbb{E} \left(\int_{N\omega}^{n\omega} \|T(t+s)\| \|F((n+p)\omega - s) - F(n\omega - s)\| ds \right)^2 \\ &\leq 4\mathbb{E} \left(\int_{N\omega}^{n\omega} M e^{-a(t+s)} \|F((n+p)\omega - s) - F(n\omega - s)\| ds \right)^2 \\ &= 4\mathbb{E} \left(\int_{N\omega}^{n\omega} M e^{-\frac{a}{2}(t+s)} e^{-\frac{a}{2}(t+s)} \|F((n+p)\omega - s) - F(n\omega - s)\| ds \right)^2. \end{aligned}$$

Again, using Cauchy Schwarz inequality, it follows that

$$\begin{aligned} I_4 &\leq 4 \int_{N\omega}^{n\omega} M^2 e^{-a(t+s)} ds \int_{N\omega}^{n\omega} e^{-a(t+s)} \mathbb{E}\|F((n+p)\omega - s) - F(n\omega - s)\|^2 ds \\ &\leq 4 \int_{N\omega}^{+\infty} M^2 e^{-as} ds \int_{N\omega}^{+\infty} e^{-as} \mathbb{E}\|F((n+p)\omega - s) - F(n\omega - s)\|^2 ds. \end{aligned}$$

Since $\mathbb{E}\|F((n+p)\omega - s) - F(n\omega - s)\|^2 \leq 2K$, we obtain

$$I_4 \leq \frac{8M^2K}{a^2} e^{-2aN\omega}$$

and hence $I_4 \leq \epsilon$.

This shows that $(X_n(t))_{n \geq 1}, t \geq 0$, is a Cauchy sequence in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$. □

Lemma 3.2 *Let F be a square mean ω -periodic limit process in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ such that*

$$\lim_{n \rightarrow +\infty} \mathbb{E}\|F(t + n\omega) - \tilde{F}(t)\|^2 = 0$$

for all $t \geq 0$. Define $V(t) = \int_0^t T(t-s)F(s)ds$. Under Assumption (H1), we have

$$\lim_{n \rightarrow +\infty} \mathbb{E}\|V(t + n\omega) - V^*(t)\|^2 = 0$$

uniformly on $t \geq 0$, where

$$V^*(t) = U(t) + \int_0^t T(t-s)\tilde{F}(s)ds.$$

Proof Let us rewrite

$$\begin{aligned} V(t+n\omega) &= \int_0^{t+n\omega} T(t+n\omega-s)F(s)ds \\ &= \int_{-n\omega}^t T(t-s)F(s+n\omega)ds \\ &= \int_{-n\omega}^0 T(t-s)F(s+n\omega)ds + \int_0^t T(t-s)F(s+n\omega)ds \\ &= \int_0^{n\omega} T(t+s)F(s+n\omega)ds + \int_0^t T(t-s)F(s+n\omega)ds \\ &= X_n(t, n) + I(t, n). \end{aligned}$$

We have

$$\begin{aligned} \mathbb{E}\|V(t+n\omega) - V^*(t)\|^2 &= \mathbb{E}\left\|X_n(t) + I(t, n) - U(t) - \int_0^t T(t-s)\tilde{F}(s)ds\right\|^2 \\ &\leq 2\mathbb{E}\|X_n(t) - U(t)\|^2 \\ &\quad + 2\mathbb{E}\left\|I(t, n) - \int_0^t T(t-s)\tilde{F}(s)ds\right\|^2. \end{aligned}$$

Using Lemma 3.1, it follows that

$$\mathbb{E}\|X_n(t) - U(t)\|^2 \rightarrow 0$$

for all $t \geq 0$.

Note that for $m\omega \leq t < (m+1)\omega$; $m \in \mathbb{N}$, one has

$$\begin{aligned} &\mathbb{E}\left\|I(t, n) - \int_0^t T(t-s)\tilde{F}(s)ds\right\|^2 \\ &= \mathbb{E}\left\|\int_0^t T(t-s)\left(F(s+n\omega) - \tilde{F}(s)\right)ds\right\|^2 \\ &\leq \mathbb{E}\left(\int_0^t \|T(t-s)\| \|F(s+n\omega) - \tilde{F}(s)\| ds\right)^2 \\ &\leq \mathbb{E}\left(\int_0^t Me^{-a(t-s)} \|F(s+n\omega) - \tilde{F}(s)\| ds\right)^2 \\ &\leq 2\mathbb{E}\left(\int_0^{m\omega} Me^{-a(t-s)} \|F(s+n\omega) - \tilde{F}(s)\| ds\right)^2 \end{aligned}$$

$$+ 2\mathbb{E} \left(\int_{m\omega}^t M e^{-a(t-s)} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2.$$

But firstly,

$$\begin{aligned} & 2\mathbb{E} \left(\int_0^{m\omega} M e^{-a(t-s)} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2 \\ & \leq 2M^2 \mathbb{E} \left(\sum_{k=0}^{m-1} \int_{k\omega}^{(k+1)\omega} e^{-a(t-(k+1)\omega)} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2 \\ & = 2M^2 \mathbb{E} \left(\sum_{k=0}^{m-1} \int_0^\omega e^{-a(t-(k+1)\omega)} \|F(s + (n+k)\omega) - \tilde{F}(s+k\omega)\| ds \right)^2 \\ & \leq 2M^2 \int_0^\omega \left(\sum_{k=0}^{m-1} e^{-a(t-(k+1)\omega)} \right)^2 ds \int_0^\omega \mathbb{E} \|F(s + (n+k)\omega) - \tilde{F}(s+k\omega)\|^2 ds. \end{aligned}$$

Since

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|F(s + (n+k)\omega) - \tilde{F}(s+k\omega)\|^2 = 0$$

for $s \in [0, \omega]$ and the fact that

$$\mathbb{E} \|F(s + (n+k)\omega) - \tilde{F}(s+k\omega)\|^2 \leq 2K,$$

it follows by Lebesgue's dominated convergence theorem that :

$$\lim_{n \rightarrow +\infty} \int_0^\omega \mathbb{E} \|F(s + (n+k)\omega) - \tilde{F}(s+k\omega)\|^2 ds = 0.$$

Note that

$$\sum_{k=0}^{m-1} e^{-a(t-(k+1)\omega)} \leq \frac{1}{1 - e^{-a\omega}} \quad \text{uniformly in } t,$$

therefore

$$2\mathbb{E} \left(\int_0^{m\omega} M e^{-a(t-s)} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2 \leq \frac{2M^2\omega}{(1 - e^{-a\omega})^2} \epsilon.$$

On the other hand,

$$2\mathbb{E} \left(\int_{m\omega}^t M e^{-a(t-s)} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2$$

$$\begin{aligned} &\leq 2M^2 \mathbb{E} \left(\int_{m\omega}^{(m+1)\omega} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2 \\ &= 2M^2 \mathbb{E} \left(\int_0^\omega \|F(s + (n + m)\omega) - \tilde{F}(s + m\omega)\| ds \right)^2 \\ &\leq 2M^2 \omega \int_0^\omega \mathbb{E} \|F(s + (n + m)\omega) - \tilde{F}(s + m\omega)\|^2 ds. \end{aligned}$$

But,

$$\begin{aligned} \mathbb{E} \|F(s + (n + m)\omega) - \tilde{F}(s + m\omega)\|^2 &\leq 2\mathbb{E} \|F(s + (n + m)\omega) - \tilde{F}(s)\|^2 \\ &\quad + 2\mathbb{E} \|\tilde{F}(s + m\omega) - \tilde{F}(s)\|^2 \end{aligned}$$

so that

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|F(s + (n + m)\omega) - \tilde{F}(s + m\omega)\|^2 = 0.$$

Again, using the Lebesgue's dominated convergence theorem, we have

$$\lim_{n \rightarrow +\infty} \int_0^\omega \mathbb{E} \|F(s + (n + m)\omega) - \tilde{F}(s + m\omega)\|^2 ds = 0,$$

and hence

$$\lim_{n \rightarrow +\infty} 2\mathbb{E} \left(\int_{m\omega}^t M e^{-a(t-s)} \|F(s + n\omega) - \tilde{F}(s)\| ds \right)^2 = 0.$$

Thus,

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left\| I(t, n) - \int_0^t T(t-s) \tilde{F}(s) ds \right\|^2 = 0$$

for all $t \geq 0$.

Therefore,

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|V(t + n\omega) - V^*(t)\|^2 = 0$$

uniformly on $t \geq 0$ for some stochastic process $V^*(t) : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$. □

Lemma 3.3 *Let G be a square mean ω -periodic limit process in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ and $B(t)$ a two-sided standard one-dimensional Brownian motion.*

Under Assumption (H1), the sequence of stochastic process $(Y_n(t))_{n \geq 1}$, $t \geq 0$ defined by

$$Y_n(t) = \int_{-n\omega}^0 T(t-s) G(s + n\omega) dB(s)$$

is a Cauchy sequence in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ for all $t \geq 0$.

We will denote by $U^* = (U^*(t))_{t \geq 0}$ the limit process of $(Y_n(t))_{n \geq 1}$, $t \geq 0$, in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Proof We have

$$\begin{aligned} \mathbb{E} \|Y_{n+p}(t) - Y_n(t)\|^2 &= \mathbb{E} \left\| \int_{-(n+p)\omega}^0 T(t-s)G((n+p)\omega + s)dB(s) \right. \\ &\quad \left. - \int_{-n\omega}^0 T(t-s)G(n\omega + s)dB(s) \right\|^2 \\ &\leq 2\mathbb{E} \left\| \int_{-(n+p)\omega}^{-n\omega} T(t-s)G((n+p)\omega + s)dB(s) \right\|^2 \\ &\quad + 2\mathbb{E} \left\| \int_{-n\omega}^0 T(t-s)(G((n+p)\omega + s) - G(n\omega + s))dB(s) \right\|^2 \\ &\leq 2\mathbb{E} \left(\int_{-(n+p)\omega}^{-n\omega} \|T(t-s)\| \|G((n+p)\omega + s)\|dB(s) \right)^2 \\ &\quad + 2\mathbb{E} \left(\int_{-n\omega}^0 \|T(t-s)\| \|G((n+p)\omega + s) - G(n\omega + s)\|dB(s) \right)^2 \\ &\leq 2\mathbb{E} \int_{-(n+p)\omega}^{-n\omega} \|T(t-s)\|^2 \|G((n+p)\omega + s)\|^2 ds \\ &\quad + 2\mathbb{E} \int_{-n\omega}^0 \|T(t-s)\|^2 \|G((n+p)\omega + s) - G(n\omega + s)\|^2 ds \\ &\leq 2\mathbb{E} \int_{n\omega}^{(n+p)\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s)\|^2 ds \\ &\quad + 2\mathbb{E} \int_0^{n\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\ &\leq J_1(t, n, p) + J_2(t, n, p) \end{aligned}$$

where

$$\begin{aligned} J_1(t, n, p) &= 2\mathbb{E} \int_{n\omega}^{(n+p)\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s)\|^2 ds \\ J_2(t, n, p) &= 2\mathbb{E} \int_0^{n\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds. \end{aligned}$$

Estimates of $J_1(t, n, p)$.

$$\begin{aligned} J_1(t, n, p) &= 2\mathbb{E} \int_{n\omega}^{(n+p)\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s)\|^2 ds \\ &\leq 2\mathbb{E} \int_{n\omega}^{(n+p)\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s)\|^2 ds \end{aligned}$$

$$\begin{aligned}
 &\leq 2 \int_{n\omega}^{(n+p)\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s)\|^2 ds \\
 &\leq 2M^2 K \int_{n\omega}^{(n+p)\omega} e^{-2as} ds \\
 &\leq 2M^2 K \int_{n\omega}^{+\infty} e^{-2as} ds \\
 &\leq \frac{2M^2 K}{2a} e^{-2an\omega}
 \end{aligned}$$

Now, we consider the integers N_1 and N_2 such that

$$\begin{aligned}
 \frac{M^2 K}{a} e^{-2an\omega} &\leq \epsilon \quad \forall n \geq N_1 \\
 \frac{4M^2 K}{a} e^{-2an\omega} &\leq \epsilon \quad \forall n \geq N_2
 \end{aligned}$$

and set $N = \max(N_1, N_2)$. For $n \geq N$, we have :

$$\begin{aligned}
 J_2(t, n, p) &= 2\mathbb{E} \int_0^{n\omega} \|T(t+s)\|^2 \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\
 &\leq 2 \int_0^{n\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\
 &\leq 2 \int_0^{N\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\
 &\quad + 2 \int_{N\omega}^{n\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\
 &\leq J_3(t, n, p) + J_4(t, n, p)
 \end{aligned}$$

where

$$\begin{aligned}
 J_3(t, n, p) &= 2 \int_0^{N\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\
 J_4(t, n, p) &= 2 \int_{N\omega}^{n\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds.
 \end{aligned}$$

Since G is a square mean ω periodic limit process, then

$$\lim_{k \rightarrow +\infty} \mathbb{E} \|G(t+k\omega) - \tilde{G}(t)\|^2 = 0, \quad \forall t \geq 0$$

is well defined for each $t \geq 0$ when $k \in \mathbb{N}$ for some stochastic process $\tilde{G} : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$.

Now,

$$\begin{aligned} J_3(t, n, p) &= 2 \int_0^{N\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds \\ &= 2 \int_0^{N\omega} M^2 e^{-2a(t-s+N\omega)} \mathbb{E} \|G((n-N+p)\omega + s) - G((n-N)\omega + s)\|^2 ds \\ &\leq 4 \int_0^{N\omega} M^2 e^{-2a(t-s+N\omega)} \mathbb{E} \|G((n-N+p)\omega + s) - \tilde{G}(s)\|^2 ds \\ &\quad + 4 \int_0^{N\omega} M^2 e^{-2a(t-s+N\omega)} \mathbb{E} \|G((n-N)\omega + s) - \tilde{G}(s)\|^2 ds \end{aligned}$$

Since

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \|G((n-N+p)\omega + s) - \tilde{G}(s)\|^2 &= 0 \\ \lim_{n \rightarrow \infty} \mathbb{E} \|G((n-N)\omega + s) - \tilde{G}(s)\|^2 &= 0 \end{aligned}$$

and using the fact that $\max\{\mathbb{E} \|G((n-N+p)\omega + s) - \tilde{G}(s)\|^2, \mathbb{E} \|G((n-N)\omega + s) - \tilde{G}(s)\|^2\} \leq 2K$, it follows by Lebesgue's dominated convergence theorem that

$$\lim_{n \rightarrow \infty} J_3(t, n, p) = 0$$

for all $t \geq 0$.

On the other hand,

$$J_4(t, n, p) = 2 \int_{N\omega}^{n\omega} M^2 e^{-2a(t+s)} \mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 ds.$$

Since $\mathbb{E} \|G((n+p)\omega - s) - G(n\omega - s)\|^2 \leq 2K$, we get

$$\begin{aligned} J_4(t, n, p) &\leq 4M^2 K \int_{N\omega}^{+\infty} M^2 e^{-2as} ds \\ &\leq \frac{4M^2 K}{a} e^{-2aN\omega} \\ &\leq \frac{4M^2 K}{a} e^{-2aN_2\omega} \end{aligned}$$

so that

$$J_4(t, n, p) \leq \epsilon.$$

This shows that $(Y_n(t))_{n \geq 1}, t \geq 0$, is a Cauchy sequence in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ □

Lemma 3.4 *Let G be square mean ω -periodic limit in $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$ such that*

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|G(t + n\omega) - \tilde{G}(t)\|^2 = 0$$

for all $t \geq 0$. Define

$$H(t) = \int_0^t T(t-s)G(s)dB(s)$$

Under Assumption (H1), we have

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|H(t + n\omega) - H^*(t)\|^2 = 0$$

uniformly on $t \geq 0$ where

$$H^*(t) = U^*(t) + \int_0^t T(t-s)\tilde{G}(s)dB(s).$$

Proof Let us rewrite,

$$\begin{aligned} H(t + n\omega) &= \int_0^{t+n\omega} T(t + n\omega - s)G(s)dB(s) \\ &= \int_{-n\omega}^t T(t-s)G(s+n\omega)dB(s+n\omega). \end{aligned}$$

Let $\tilde{B}(s) = B(s + n\omega) - B(n\omega)$ for each $s \in \mathbb{R}$. By the weak Markov property \tilde{B} is also a two sided Brownian motion and has the same distribution as B . Moreover, $\{\tilde{B}(s), s \in \mathbb{R}\}$ is a two-sided Brownian motion independent of $B(n\omega)$. Thus,

$$\begin{aligned} H(t + n\omega) &= \int_{-n\omega}^t T(t-s)G(s+n\omega)d\tilde{B}(s) \\ &= \int_{-n\omega}^0 T(t-s)G(s+n\omega)d\tilde{B}(s) + \int_0^t T(t-s)G(s+n\omega)d\tilde{B}(s) \\ &= Y_n(t) + J(t, n) \end{aligned}$$

where

$$J(t, n) = \int_0^t T(t-s)G(s+n\omega)d\tilde{B}(s) = \int_0^t T(t-s)G(s+n\omega)dB(s).$$

We have

$$\begin{aligned} & \mathbb{E} \|H(t + n\omega) - H^*(t)\|^2 \\ &= \mathbb{E} \left\| Y_n(t) + J(t, n) - U^*(t) - \int_0^t T(t-s)\tilde{G}(s)ds \right\|^2 \\ &\leq \mathbb{E} \|Y_n(t) - U^*(t)\|^2 \\ &\quad + \mathbb{E} \left\| J(t, n) - \int_0^t T(t-s)\tilde{G}(s)ds \right\|^2 \end{aligned}$$

Using Lemma 3.3, it follows that

$$\mathbb{E} \|Y_n(t) - U^*(t)\|^2 \rightarrow 0$$

for all $t \geq 0$, when $n \rightarrow +\infty$.

For $m\omega \leq t < (m + 1)\omega$; $m \in \mathbb{N}$, one has

$$\begin{aligned} & \mathbb{E} \left\| J(t, n) - \int_0^t T(t-s)\tilde{G}(s)dB(s) \right\|^2 \\ &= \mathbb{E} \left\| \int_0^t T(t-s) \left(G(s + n\omega) - \tilde{G}(s) \right) dB(s) \right\|^2 \\ &\leq \mathbb{E} \left(\int_0^t \|T(t-s)\| \|G(s + n\omega) - \tilde{G}(s)\| dB(s) \right)^2 \\ &\leq \mathbb{E} \left(\int_0^t M e^{-a(t-s)} \|G(s + n\omega) - \tilde{G}(s)\| dB(s) \right)^2 \\ &= \int_0^t M^2 e^{-2a(t-s)} \mathbb{E} \|G(s + n\omega) - \tilde{G}(s)\|^2 ds \\ &\leq \int_0^{m\omega} M^2 e^{-2a(t-s)} \mathbb{E} \|G(s + n\omega) - \tilde{G}(s)\|^2 ds \\ &\quad + \int_{m\omega}^t M^2 e^{-2a(t-s)} \mathbb{E} \|G(s + n\omega) - \tilde{G}(s)\|^2 ds. \end{aligned}$$

Now,

$$\begin{aligned} & \int_0^{m\omega} M^2 e^{-2a(t-s)} \mathbb{E} \|G(s + n\omega) - \tilde{G}(s)\|^2 ds \\ &\leq M^2 \sum_{k=0}^{m-1} \int_{k\omega}^{(k+1)\omega} e^{-2a(t-(k+1)\omega)} \mathbb{E} \|G(s + n\omega) - \tilde{G}(s)\|^2 ds \\ &= M^2 \int_0^\omega \sum_{k=0}^{m-1} e^{-2a(t-(k+1)\omega)} \mathbb{E} \|G(s + (n+k)\omega) - \tilde{G}(s+k\omega)\|^2 ds \end{aligned}$$

$$\begin{aligned} &\leq 2M^2 \int_0^\omega \sum_{k=0}^{m-1} e^{-2a(t-(k+1)\omega)} \mathbb{E} \|G(s + (n+k)\omega) - \tilde{G}(s)\|^2 ds \\ &\quad + 2M^2 \int_0^\omega \sum_{k=0}^{m-1} e^{-2a(t-(k+1)\omega)} \mathbb{E} \|\tilde{G}(s+k\omega) - \tilde{G}(s)\|^2 ds \\ &= 2M^2 \int_0^\omega \sum_{k=0}^{m-1} e^{-2a(t-(k+1)\omega)} \mathbb{E} \|G(s + (n+k)\omega) - \tilde{G}(s)\|^2 ds \end{aligned}$$

because

$$\mathbb{E} \|\tilde{G}(s+k\omega) - \tilde{G}(s)\|^2 = 0 \quad \forall k \geq 0.$$

Since

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|G(s + (n+k)\omega) - \tilde{G}(s)\|^2 = 0,$$

for $s \in [0, \omega]$, $k \geq 0$ with $\mathbb{E} \|G(s + (n+k)\omega) - \tilde{G}(s)\|^2 \leq 2K$ and using the fact that

$$\sum_{k=0}^{m-1} e^{-2a(t-(k+1)\omega)} \leq \frac{1}{1 - e^{-2a\omega}} \quad \forall t \geq 0,$$

it follows by Lebesgue's dominated convergence theorem that :

$$\lim_{n \rightarrow +\infty} \int_0^\omega \sum_{k=0}^{m-1} e^{-2a(t-(k+1)\omega)} \mathbb{E} \|G(s + (n+k)\omega) - \tilde{G}(s)\|^2 ds = 0$$

uniformly on $t \geq 0$.

On the other hand,

$$\begin{aligned} &\int_{m\omega}^t M^2 e^{-2a(t-s)} \mathbb{E} \|G(s+n\omega) - \tilde{G}(s)\|^2 ds \\ &\leq \int_{m\omega}^{(m+1)\omega} M^2 e^{-2a(t-s)} \mathbb{E} \|G(s+n\omega) - \tilde{G}(s)\|^2 ds \\ &= \int_0^\omega M^2 e^{-2a(t-s-m\omega)} \mathbb{E} \|G(s+(n+m)\omega) - \tilde{G}(s+m\omega)\|^2 ds \\ &\leq M^2 \int_0^\omega \mathbb{E} \|G(s+(n+m)\omega) - \tilde{G}(s+m\omega)\|^2 ds \\ &\leq 2M^2 \int_0^\omega \mathbb{E} \|G(s+(n+m)\omega) - \tilde{G}(s)\|^2 ds \\ &\quad + 2M^2 \int_0^\omega \mathbb{E} \|\tilde{G}(s+m\omega) - \tilde{G}(s)\|^2 ds \end{aligned}$$

$$= 2M^2 \int_0^\omega \mathbb{E} \|G(s + (n + m)\omega) - \tilde{G}(s)\|^2 ds$$

because

$$\mathbb{E} \|\tilde{G}(s + m\omega) - \tilde{G}(s)\|^2 = 0 \quad \forall m \geq 0.$$

Since

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|G(s + (n + m)\omega) - \tilde{G}(s)\|^2 = 0,$$

for $s \in [0, \omega]$, $m \geq 0$ and

$$\mathbb{E} \|G(s + (n + m)\omega) - \tilde{G}(s)\|^2 \leq 2K,$$

again by the Lebesgue dominated convergence theorem, we have :

$$\lim_{n \rightarrow +\infty} 2M^2 \int_0^\omega \mathbb{E} \|G(s + (n + m)\omega) - \tilde{G}(s)\|^2 ds = 0$$

so that

$$\int_{m\omega}^t M^2 e^{-2a(t-s)} \mathbb{E} \|G(s + n\omega) - \tilde{G}(s)\|^2 ds.$$

In view of the above, it follows that

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left\| J(t, n) - \int_0^t T(t-s) \tilde{G}(s) dB(s) \right\|^2 = 0$$

uniformly on $t \geq 0$.

Therefore,

$$\lim_{n \rightarrow +\infty} \mathbb{E} \|H(t + n\omega) - H^*(t)\|^2 = 0$$

uniformly in $t \geq 0$ for the stochastic process $H^*(t) : \mathbb{R}_+ \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ defined as above. □

Now, we can establish the main result of this section.

Theorem 3.5 *Let $f, g : \mathbb{R}_+ \times \mathbb{L}^2(\mathbb{P}, \mathbb{H}) \rightarrow \mathbb{L}^2(\mathbb{P}, \mathbb{H})$ be square mean ω periodic limit processes in $t \geq 0$ uniformly in X for bounded subsets of $\mathbb{L}^2(\mathbb{P}, \mathbb{H})$. Assume that f, g satisfies a Lipschitz condition, uniformly in $t \geq 0$: that is, there exist constants $L_f > 0$ and $L_g > 0$ such that*

$$\mathbb{E} \|f(t, X) - f(t, Y)\|^2 \leq L_f \mathbb{E} \|X - Y\|^2 \quad \forall t \geq 0, \forall X, Y \in \mathbb{L}^2(\mathbb{P}, \mathbb{H})$$

$$\mathbb{E} \|g(t, X) - g(t, Y)\|^2 \leq L_g \mathbb{E} \|X - Y\|^2 \quad \forall t \geq 0, \forall X, Y \in \mathbb{L}^2(\mathbb{P}, \mathbb{H}).$$

If

$$2M^2 \left(L_f \frac{1}{a^2} + L_g \frac{1}{a} \right) < 1$$

then there is a unique square mean asymptotically ω -periodic mild solution of problem (12).

Proof We define the continuous operator Γ on the Banach space $AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ by

$$(\Gamma X)(t) = T(t)c_0 + \int_0^t T(t-s)f(s, X(s))ds + \int_0^t T(t-s)g(s, X(s))dB(s).$$

Note that $T(t)c_0$ is in $C_0(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H})) \subseteq AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$

Now, we denote $F(s) = f(s, X(s))$, $G(s) = g(s, X(s))$

In view of Theorem 2.5 if $X \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$, then $F, G \in P_\omega L(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$.

Applying Lemma 3.2, Lemma 3.4 and Theorem 2.4, it follows that

$$\int_0^t T(t-s)f(s, X(s))ds \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$$

and

$$\int_0^t T(t-s)g(s, X(s))dB(s) \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H})).$$

Hence, the operator Γ maps the space $AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ into itself.

Finally, for any $X, Y \in AP_\omega(\mathbb{R}_+, \mathbb{L}^2(\mathbb{P}, \mathbb{H}))$ we have

$$\begin{aligned} & \mathbb{E} \|\Gamma X(t) - \Gamma Y(t)\|^2 \\ & \leq 2M^2 \left(\int_0^t e^{-a(t-s)} ds \right) \mathbb{E} \int_0^t e^{-a(t-s)} \|f(s, X(s)) - f(s, Y(s))\|^2 ds \\ & \quad + 2M^2 \mathbb{E} \int_0^t e^{-2a(t-s)} \|g(s, X(s)) - g(s, Y(s))\|^2 ds \\ & \leq 2M^2 L_f \sup_{s \geq 0} \mathbb{E} \|X(s) - Y(s)\|^2 \left(\int_0^t e^{-a(t-s)} ds \right)^2 \\ & \quad + 2M^2 L_g \sup_{s \geq 0} \mathbb{E} \|X(s) - Y(s)\|^2 \int_0^t e^{-2a(t-s)} ds \\ & \leq 2M^2 \left(L_f \frac{1}{a^2} + L_g \frac{1}{a} \right) \sup_{s \geq 0} \mathbb{E} \|X(s) - Y(s)\|^2. \end{aligned}$$

This implies that

$$\|\Gamma X - \Gamma Y\|_\infty^2 \leq 2M^2 \left(L_f \frac{1}{a^2} + L_g \frac{1}{a} \right) \|X - Y\|_\infty^2.$$

Consequently, if $2M^2 \left(L_f \frac{1}{a^2} + L_g \frac{1}{a} \right) < 1$, then Γ is a contraction mapping.

The proof is completed by using the well-known Banach fixed-point theorem. \square

4 An Illustrative Example

In order to illustrate usefulness of the theoretical results established in the preceding section, we consider the following one-dimensional stochastic heat equation with Dirichlet boundary conditions :

$$\begin{cases} du(t, x) = \frac{\partial^2 u(t, x)}{\partial x^2} dt + f(t, u(t, x))dt + g(t, u(t, x))dB(t) \\ u(t, 0) = u(t, 1) = 0, t \in \mathbb{R}^+, \\ u(0, x) = h(x), x \in [0, 1]. \end{cases} \quad (13)$$

where $B(t)$ is a two-sided standard one-dimensional Brownian motion defined on the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$, $h \in L^2[0, 1]$ and the functions f and g are defined as

$$f(t, u(t, x)) = u(t, x)\psi(t) \quad \text{and} \quad g(t, u(t, x)) = u(t, x)\phi(t),$$

where ψ and ϕ are ω -periodic limit (function) deterministic processes. Clearly, both the f and g satisfy the Lipschitz conditions with $L_f = \|\psi\|_\infty$ and $L_g = \|\phi\|_\infty$. For instance, $\psi(t)$ and $\phi(t)$ can be chosen to be equal to the following 2-periodic limit (function) deterministic process (see [18]) given by

$$a_{\{k_n\}}(t) = \begin{cases} 1, & t = 2n - 1, n \in \mathbb{N} \\ 0, & t \in \{0, 2\} \cup \{2n + 1 - k_n\} \cup \{2n + 1 + k_n\} \\ \text{linear,} & \text{in between.} \end{cases} \quad (14)$$

where $\{k_n\} \subset]0, 1[$ such that $k_n > k_{n+1}$, $k_n \rightarrow 0$ as $n \rightarrow +\infty$.

Note that if we define $b : \mathbb{R}_+ \rightarrow \mathbb{R}$ by

$$b(t) = \begin{cases} 1, & t = 2n - 1, n \in \mathbb{N} \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

then we have $b(t) = \lim_{m \rightarrow +\infty} a_{\{k_n\}}(t + 2m)$ so $a_{\{k_n\}}$ is a 2-periodic limit (function) deterministic process.

Define

$$\mathcal{D}(A) = \{v \text{ continuous}/v'(r) \text{ absolutely continuous on } [0, 1], v''(r) \in L^2[0, 1] \text{ and } v(0) = v(1) = 0\}$$

$$Av = v'' \text{ for all } v \in \mathcal{D}(A).$$

Let $\phi_n(t) = \sqrt{2} \sin(n\pi t)$ for all $n \in \mathbb{N}$. ϕ_n are eigenfunctions of the operator $(A, \mathcal{D}(A))$ with eigenvalues $\lambda_n = -n^2$. Then, A generates a C_0 semigroup $(T(t))$ of the form

$$T(t)\phi = \sum_{n=1}^{\infty} e^{-n^2\pi^2 t} \langle \phi, \phi_n \rangle \phi_n, \quad \forall \phi \in L^2[0, 1]$$

and

$$\|T(t)\| \leq e^{-\pi^2 t}, \quad \text{for all } t \geq 0$$

Thus, $M = 1$ and $a = \pi^2$.

Equation (12) is of the form

$$\begin{cases} dy(t) = Ay(t)dt + f(t, y(t))dt + g(t, y(t))dB(t), \\ y(0) = c_0. \end{cases}$$

By using Theorem 3.5, we claim that

Theorem 4.1 *If $\|\psi\|_{\infty} + \|\phi\|_{\infty} \pi^2 < \pi^4/2$, then equation (13) admits a unique square mean asymptotically ω -periodic mild solution.*

Acknowledgements We thank the anonymous reviewers for their careful reading and many insightful comments and suggestions.

Compliance with Ethical Standards

Conflict of interests The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Corduneanu, C.: *Almost Periodic Oscillations and Waves*. Springer, New York (2009)
2. Kundert, K.S., Sorkin, G.B., Sangiovanni-Vincentelli, A.: Applying harmonic balance to almost periodic circuits. *IEEE Trans. Microw. Theory Tech.* **36**(2), 366–378 (1988)
3. Ahmad, S.: On almost periodic solutions of the competing species problems. *Proc. Am. Math. Soc.* **102**(4), 855–861 (1988)
4. Bezandry, P., Diagana, T.: Square-mean almost periodic solutions nonautonomous stochastic differential equations. *Electron. J. Differ. Equ.* **2007**(117), 1–10 (2007)
5. Chang, Y.K., Zhao, Z.H., N'Guérékata, G.M.: A new composition theorem for square-mean almost automorphic functions and applications to stochastic differential equations. *Nonlinear Anal.: Theory Methods Appl.* **75**(6), 2210–2219 (2011)

6. Bezandry, P., Diagana, T.: Existence of square-mean almost periodic solutions to some stochastic hyperbolic differential equations with infinite delay. *Commun. Math. Anal.* **8**(2), 103–124 (2010)
7. Manou-Abi, S.M., Dimbour, W.: S -Asymptotically ω -periodic solutions in the p -th mean for a stochastic evolution equation driven by Q -Brownian motion. *Adv. Sci. Technol. Eng. Syst. J.* **2**(5), 124–133 (2017)
8. Cao, J., Yang, Q., Huang, Z., Liu, Q.: Asymptotically almost periodic solutions of stochastic functional differential equations. *Appl. Math. Comput.* **218**, 1499–1511 (2011)
9. Chang, Y.K., Zhao, Z.H., N'Guérékata, G.M.: Square mean almost automorphic mild solutions to non-autonomous stochastic differential equations in Hilbert spaces. *Comput. Math. Appl.* **61**, 384–391 (2011)
10. Cuevas, C., de Souza, J.C.: S -Asymptotically ω -periodic solutions of semilinear fractional integro-differential equations. *Appl. Math. Lett.* **22**, 865–870 (2009)
11. Dimbour, W., Manou-Abi, S.M.: Asymptotically ω -periodic solution for an evolution differential equation via ω -periodic limit functions. *Int. J. Pure Appl. Math.* **113**(1), 59–71 (2017)
12. Dimbour, W., Manou-Abi, S.M.: Asymptotically ω -periodic functions in the Stepanov sense and its application for an advanced differential equation with piecewise constant argument in a Banach space. *Mediterr. J. Math.* **15**, 25 (2018)
13. Liu, Z., Sun, K.: Almost automorphic solutions to SDE driven by Levy noise. *J. Funct. Anal.* **266**(3), 1115–1149 (2014)
14. Xia, Z.: Almost automorphic solutions semilinear stochastic hyperbolic differential equations in intermediate space. *Kodai Math. J.* **40**(3), 492–517 (2017)
15. Xia, Z., Wang, D.: Measure pseudo almost periodic mild solutions of stochastic functional differential equations with Lévy noise. *J. Nonlinear Convex Anal.* **18**(5), 847–858 (2017)
16. Xie, R., Zhang, C.: Criteria of asymptotic ω periodicity and their applications in a class of fractional differential equations. *Adv. Differ. Equ.* **2015**, 68 (2015)
17. Zhang, M., Zong, G.: Almost periodic solutions for stochastic differential equations driven by G-Brownian motion. *Commun. Stat. Theory Methods* **44**(11), 2371–2384 (2015)
18. Xie, R., Zhang, C.: Space of ω periodic limit functions and its applications to an abstract cauchy problem. *J. Funct. Spaces* **2015**, 10 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bibliographie

- [1] Cont Rama (2009). La statistique face aux événements rares. *Pour la science*, (385), 116-123.
- [2] Hill B.M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163-1174.
- [3] C. Cuevas, J. de Souza. *Existence of S -asymptotically ω -periodic solutions for fractional order functional integro-differential equations with infinite delay*. *Nonlinear Analysis*, 72, 2010.
- [4] W. Dimbour, V. Valmorin. *Asymptotically antiperiodic solutions for a nonlinear differential equation with piecewise constant argument in a Banach space*. *Applied Mathematics*, 7 (2016).
- [5] W. Dimbour, S. Manou-Abi. *S -asymptotically ω -periodic solution for a nonlinear differential equation with piecewise constant argument via S -asymptotically ω -periodic functions in the Stepanov sense*. *International J. of Pure and Appl. Math.*
- [6] W. Dimbour, G.Mophou and G.M. N'Guérékata. *S asymptotically ω -periodic solution for partial differential equations with finite delay*. *Electron.J.Differ.Equa.* 2011, 1-12, 2011.
- [7] H. R. Henríquez, M. Pierri and P. Táboas. *On S asymptotically ω -periodic function on Banach spaces and applications*. *J. Math. Anal Appl.* 343, 1119-1130, 2008.
- [8] H. R. Henríquez, M. Pierri and P. Táboas. *Existence of S -asymptotically ω -periodic solutions for abstract neutral equations*. *Bull. Aust. Math.Soc.* 78, 365-382, 2008.
- [9] G.M. N'Guérékata, V. Valmorin. *Antiperiodic solutions of semilinear integrodifferential equations in Banach spaces*. *Applied Mathematics and Computation*, 218, 2012.
- [10] S. Nicola, M. Pierri. *A note on S -asymptotically periodic functions*. *Nonlinear Analysis, Real World Application*, 10 (2009).
- [11] M. Pierri. *On S -Asymptotically ω -periodic functions and applications*. *Nonlinear Anal.* 75, 651-661, 2012.
- [12] Rong-Hua, He. *Stepanov-like pseudo-almost automorphic mild solutions for some abstract differential equations.*, *Advances in Fixed Point Theory*, 2(3),258-272, 2012

- [13] J.Wiener, *A Second-Order delay differential equation with multiple Periodic solutions*, Journal of Mathematical Analysis and Application, 229 (1999) 6596-676.
- [14] J.Wiener, L.Debnath *Boundary Value Problems for the diffusion equation with piecewise continuous time delay*, Internat.J.Math.and Math.Sci., Vol.20 (1997) 187-195.
- [15] J.Wiener, L.Debnath *A survey of partial differential equations with piecewise continuous arguments*, Internat.J.Math.and Math.Sci., Vol.18, No.2 (1995) 209-228.
- [16] J.Wiener, V.Lakshmikantham, *Excitability of a second-order delay differential equation*, Nonlinear Analysis, Vol.38 (1999) 1-11.
- [17] J.Wiener, *Generalized solutions of functional differential equations*, World Scientific (1999).
- [18] R. Xie and C. Zhang. *Criteria of asymptotic ω -periodicity and their applications in a class of fractional differential equations. Advances in Difference Equations, 2015.*
- [19] Z.Xia. *Asymptotically periodic of semilinear fractional integro-differential equations. Advances in Difference Equations, 1-19, 2014 .*
- [20] Z.Xia. *Weighted pseudo asymptotically periodic mild solutions of evolutions equations. Acta Mathematica Sinica, 31(8), 1215-1232, 2015.*
- [21] R. Xie and C. Zhang, *Space of ω periodic limit functions and its applications to an Abstract Cauchy problem, Journal of Function Spaces, 2015, ID 953540, 10 pages (2015)*
- [22] Bohr, H. (1925). *Zur theorie der fast periodischen funktionen : I. Eine verallgemeinerung der theorie der fourierreihen.* Mathematica, 45(1), 29-127.
- [23] Favard, J. (1963). *Sur certains systèmes différentiels scalaires linéaires et homogènes à coefficients presque-périodiques.* Annali di Matematica Pura ed Applicata, Series 4, 62(1), 297-316.
- [24] Schoutens, W. (2003). *Lévy Processes in Finance : Pricing Financial Derivatives.* Wiley, Hoboken, NJ, USA.
- [25] Mikosch, T., Resnick, S., Rootzen, H. and Stegeman, A. (2002). *Is network traffic approximated stable Lévy motion or fractional Brownian motion.* Ann. Appl. Probab. 12 23-68. MR1890056
- [26] Schertzer, D., Larcheveque, M., Duan, J., Yanovsky, V.V. and Lovejoy, S. (2001). *Fractional Fokker-Planck equation for nonlinear stochastic differential equations driven by non-Gaussian Lévy stable noises.* J. Math. Phys. 42 200-212. MR1808774
- [27] Ditlevsen, P.D. (1999). *Observation of α -stable noise induced millennial climate changes from an ice-core record.* Geophysical Research Letter 26 1441-1444.
- [28] Ditlevsen, P.D. (1999). *Anomalous jumping in a double-well potential.* Phys. Rev. E 60 172-179.
- [29] Yousri Slaoui. (2021). *Recursive kernel regression estimation under α -mixing data*

- [30] Hongwei Long and Lianfen Qian. Nadaraya-Watson estimator for stochastic processes driven by stable Lévy motions. *Electronic Journal of Statistics* Vol. 7 (2013) 1387–1418 ISSN : 1935-7524.
- [31] Lin Zheng Yan, Song Yu Ping and Yi Jiang Sheng. Local linear estimator for stochastic differential equations driven by α -stable Lévy motions. *Sci. China Math.* 57, 609–626 (2014). <https://doi.org/10.1007/s11425-013-4628-7>
- [32] Gilles Christ Dansou. Rapport de Stage DANSOU, Master2 de l'Université de Montpellier 2022-2023.
- [33] Mathieu Fontaine. Processus alpha-stables pour le traitement du signal. Traitement du signal et de l'image [eess.SP]. Université de Lorraine, 2019. Français. NNT : 2019LORR0037. tel-02188304
- [34] John P. Nolan. *Univariate Stable Distributions Models for Heavy Tailed Data*. Springer Series in Operations Research and Financial Engineering
- [35] Masuda, H. (2007). Ergodicity and exponential β -mixing bounds for multidimensional diffusions with jumps. *Stochastic Process. Appl.* 117 35-56. MR2287102
- [36] Hu, Y. and Long, H. (2007). Parameter estimation for OrnsteinUhlenbeck processes driven by α -stable Lévy motions. *Communication on Stochastic Analysis* 1 175-192. MR2397392
- [37] Hu, Y. and Long, H. (2009). Least square estimator for Ornstein Uhlenbeck processes driven by α -stable motions. *Stochastic Process. Appl.* 119 2465-2480. MR2532208
- [38] Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge. MR1161622
- [39] Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes*. Lecture Notes in Statistics, Vol. 110, Springer-Verlag, New York. MR1441072
- [40] Wu, W.B. (2003). Nonparametric estimation for stationary processes. Technical Report No. 536, Department of Statistics, The University of Chicago.
- [41] Wu, W.B. (2005). Nonlinear system theory : another look at dependence. *Proc. Natl. Acad. Sci. USA* 102 14150-14154. MR2172215
- [42] Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* 87 998-1004. MR1209561
- [43] Fan, J. and Zhang, C. (2003). A reexamination of diffusion estimators with applications to financial model validation. *J. Amer. Statist. Assoc.* 98 118-134. MR1965679
- [44] Spokoiny, V.G. (2000). Adaptive drift estimation for nonparametric diffusion model. *Ann. Statist.* 28 815-836. MR1792788
- [45] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London. MR1383587
- [46] Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* 9 141-142.

- [47] Wastson, G. S. (1964). Smooth regression analysis. *Sankhyi* 26 359-372.
- [48] Tsybakov, A. B. (1982a). Nonparametric signal estimation when there is incomplete information on the noise distribution. *Problemy Pereckdi Informacii* 18(2) 44-60. English translation in *Problems Inform. Transmission* 18 116-130.
- [49] Peng, L. and Yao, Q. (2004). Nonparametric regression under dependent errors with infinite variance. *Ann. Inst. Statist. Math.* 56 73-86. MR2053729.
- [50] Hall, P., Peng, L. and Yao, Q. (2002). Prediction and nonparametric estimation for time series with heavy tails. *J. Time Ser. Anal.* 23 313-331. MR1908594.
- [51] Chambers, J. M. et al. (1976). A method for simulating stable random variables. *Journal of the american statistical association*, 71(354), 340-344.
- [52] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society : Series B (Methodological)*, 53(3), 683-690.
- [53] Slaoui, Y. (2014). Bandwidth selection for recursive kernel density estimators defined by stochastic approximation method. *Journal of Probability and Statistics*, 2014.
- [54] Kogon, S. M. and Williams, D. B. (1998). Characteristic function based estimation of stable distribution parameters. *A practical guide to heavy tails : statistical techniques and applications*, 311-338.
- [55] Castillo-Barnes and al. (2020). Expectation–Maximization algorithm for finite mixture of stable distributions. *Neurocomputing*, 413, 210-216.
- [56] Salas-Gonzalez, D and al. (2010). Modelling with mixture of symmetric stable distributions using Gibbs sampling. *Signal processing*, 90(3), 774-783.
- [57] Corless, R. M. and al. (1993). Lambert’s W function in Maple. *Maple Technical Newsletter*, 9(1), 12-22.
- [58] Baddeley A. and Waagepetersen R. (2000). Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3), 329-35
- [59] de Haan, L., Mercadier, C. and Zhou, C. (2016), Adapting extreme value statistics to financial time series : dealing with bias and serial dependence, *Finance and Stochastics* 20, 321–354.
- [60] Drees H. (2000), Weighted approximations of tail processes for β -mixing random variables, *Annals of Applied Probability* 10, 1274–1301.
- [61] Dekkers, A. & de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation, *Ann. Statist.* 17 (198), 1795-1832.
- [62] Beirlant J., Dierckx G., Goegebeur Y., and Matthys G. (1999), Tail index estimation and an exponential regression model, *Extremes* 2, 177-200.
- [63] Gomes M.I., de Haan L., Rodrigues L.(2008), Tail index estimation for heavy-tailed models : accommodation of bias in weighted log-excesses. *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 70, 31–52.

- [64] Guillou A., C.-De moulin V.(2018) Extreme quantile estimation for β -mixing time series and applications. Insurance : Mathematics and Economics, 59-74.
- [65] Feuerverger A., Hall P. (1999), Estimating a tail exponent by modelling departure from a Pareto distribution, Annals of Statistics 27, 760–781.
- [66] Deme, E. H., Gardes, L., and Girard, S. (2013). On the estimation of the second order parameter for heavy-tailed distributions. Statistical Journal 3(11), 277-299.
- [67] He X, Lau EH, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. (2020) 26 :672–5. doi : 10.1038/s41591-020-0869-5
- [68] Mbaye, M. M and Manou-Abi, S. M. (2023). Existence of almost automorphic solution in distribution for a class of stochastic integro-differential equation driven by Lévy noise. The Journal of Analysis (Springer), 1-24.
- [69] Barry, M. A., Deme, E. H., Diop, A. and Manou-Abi, S. M. (2023). Improved estimators of tail index and extreme quantiles under dependence serials. Mathematical Methods of Statistics (Springer), 32(2), 133-153.
- [70] Modou Kebe, Deme, E. H., Tchilabalo Abozou Kpanzou, Manou-Abi, S. M. and Ebrima Sisawo (2023). Kernel Estimation of the Quintile Share Ratio index of Inequality for Heavy-tailed Income distributions. European Journal of Pure and Applied Mathematics, 16(4), 2509–2543.
- [71] Manou-Abi Solym M., Slaoui Yousri and Balicchi, Julien. Estimation of some epidemiological parameters with the Covid-19 data of Mayotte. Frontiers in Applied Mathematics and Statistics, 2022, vol. 8, p. 870-885.
- [72] Manou-Abi, S. M. and Dimbour, W. (2020). Asymptotically periodic solution of a stochastic differential equation. Bulletin of the Malaysian Mathematical Sciences Society (Springer), 43(1), 911-939.
- [73] Dimbour, W. and Manou-Abi, S. M. (2018). Asymptotically ω -periodic functions in the Stepanov sense and its application for an advanced differential equation with piecewise constant argument in a Banach space. Mediterranean Journal of Mathematics (Springer), 15(1), 25.
- [74] Dimbour, William. and Manou-Abi, Solym M. (2018). S-asymptotically ω -periodic solution for a nonlinear differential equation with piecewise constant argument via S-asymptotically ω -periodic functions in the Stepanov sense. J. Nonlinear Syst. Appl. Vol 7(1), 14–20.
- [75] Dimbour, William. and Manou-Abi, Solym Mawaki (2017). Asymptotically periodic solution for an evolution differential equation via periodic limits functions. Int. J. Pure Appl. Math. Vol. 113. 59-71.
- [76] Joulin, Aldéric, and Manou-Abi Solym Mawaki (2015). A note on convex ordering for stable stochastic integrals. Stochastics An International Journal of Probability and Stochastic Processes 87.4 (2015) : 592-603.

- [77] Cattiaux, P. and Manou-Abi, S.M. (2014). Limit theorems for some functionals with heavy tails of a discrete time Markov chain. *ESAIM : Probability and Statistics*, 18, 468-482.
- [78] Nguala, J. B., Manou-Abi, S., Raheiririna, A. and Slaoui, Y. (2023). Jeux et arts traditionnels, supports d'apprentissage des mathématiques. *Microscop*. Octobre 2023.
- [79] Manou-Abi, S. and Slaoui, Y. (2022). À Mayotte, une approche mathématique de la pandémie. *Microscop*. Mars 2022.
- [80] Tsilefa, Stefana Fandresena, Solym Manou-Abi, and Angelo Raheiririna. (2022). "Numerical convergence of some stochastic compartmental models in epidemiology." *CARI 2022*.
- [81] Manou-Abi, S. M., Dabo-Niang, S. and Salone, J. J. (Eds.). (2020). *Mathematical modeling of random and deterministic phenomena*. John Wiley & Sons.
- [82] Manou-Abi, S. M., Dimbour, W. and Mbaye, M. M. (2020). Existence of an Asymptotically Periodic Solution for a Stochastic Fractional Integro-differential Equation. *Mathematical Modeling of Random and Deterministic Phenomena*, 113-139.
- [83] S. M. Manou-Abi and W. Dimbour. S-asymptotically ω -periodic solutions in the p -th mean for a Stochastic Evolution Equation driven by \mathbb{Q} -Brownian motion. *Proceedings of ICAM'17. Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 5, 124-133*
- [84] Manou-Abi, S., Hachim, S. S., Dabo-Niang, S., and Nguala, J. B. . Comparative clustering methods : KL divergence, Rao distance, Bregman divergence with application to the Fani Maoré marine volcano earthquake data.
- [85] Kebe, M., Deme, E. H., Slaoui, Y. and Manou-Abi, S. M. Robust estimator of the Ruin Probability in infinite time for heavy-tailed distributions.(2023). *Statistics (Taylor & Francis)*. To appear.
- [86] Manou-Abi, S.M. Approximate solution for a class of stochastic differential equation driven by stable processes. Preprint (2023). Under revision.
- [87] Manou-Abi, S.M. Parameter Estimation for a Class of Stable Driven Stochastic Differential Equations. *Indian Journal of Probability and Statistics*, Springer (2024). To appear.
- [88] Manou-Abi, Solym, et al. "Spatio-temporal modeling and Machine Learning for mosquitoes abundance with consideration of environmental data in the island of Mayotte." Under revision.
- [89] Diouf, M. B., Deme, H., Manou-Abi, S. M. and Slaoui, Y. Box-Cox transformation on the estimation of extreme value index (EVI) and high quantiles for heavy-tailed distributions under dependence serials. *ALEA Lat. Am. J. Probab. Math. Stat. Revised*.
- [90] Hajjaji, Omar, Solym Manou-Abi, and Yousri Slaoui. "Parameter estimation for stable distributions and their mixture. (2024) *Journal of Applied Statistics*, 1-34 (Taylor & Francis).

- [91] Nguala, Jean-Berky, and Manou-Abi, Solym M. "Statistical modeling of the attitudes toward mathematics from a school survey data of Mayotte." Preprint (2024).
- [92] Paul A. L. Faye, Elodie Brunel, Thomas Claverie, Solym M. Manou-Abi and Sophie Dabo-Niang. Automatic geomorphology mapping using statistical learning algorithms. (2024) *Earth Science Informatics*, 1-18 (Springer).
- [93] Bouzalmat, Ibrahim, Benoîte de Saporta, and Solym M. Manou-Abi. "Parameter estimation for a hidden linear birth and death process with immigration." arXiv preprint arXiv :2303.00531 Preprint (2023).
- [94] Mame Birame Diouf, Hadji Deme, Solym M Manou-Abi, Yousri Slaoui (2023) Kernel estimator of extreme value index (EVI) and high quantiles for heavy-tailed distributions under dependence serials using the Box-Cox transformation. *Afr. Stat.* 18(4) : 3651-3695 (Project Euclide)
- [95] Solym Manou-Abi, Damien Devault, Sophie Dabo, Jean-Baptiste Charlier and Jean-François Desprats. Spatial modeling of the dynamic transformation of the chlordecone in soils of French West Indies from environmental factors. Preprint (2023).
- [96] Dimbour, W. (2013). Solutions presque automorphes et S asymptotiquement périodiques pour une classe d'équations d'évolution. (Doctoral dissertation, Antilles-Guyane).
- [97] Solym Manou-Abi et al. Parameter estimation of stable distributions using extreme value distributions and recursive estimators. Work in Progress.
- [98] Angelo Raherinirina, Stefana Tabera Tsilefa, T. Nirilanto and Solym Manou-Abi Bayesian inference of a spatially dependent semi-Markovian model with application to Madagascar Covid'19 data. Preprint 2024.
- [99] Solym M. Manou-Abi, Essoham Ali, Yousri Slaoui and Julien Balicchi. Estimating social contact matrices from a Zero inflated Bell model through Density Power Divergence estimation : application to a sample survey data from the island of Mayotte. Preprint 2024.
- [100] Mamadou M. Mbaye, Amadou Diop, Solym Manou-Abi and Moustapha Dieye. Approximate solutions of a stochastic partial integro-differential equations in Hilbert spaces. Work in Progress.
- [101] Stefana Tabera Tsilefa, Solym Manou-Abi, William Dimbour and Angelo Raherinirina. Approximate Bloch periodic solutions for stable driven SDE. Work in Progress.
- [102] Ibrahim Bouzalmat, Benoîte de Saporta and Solym Manou-Abi. Estimating parameters of continuous-time multi-chain hidden Markov models for infectious diseases. Preprint (2024).
- [103] Manou-Abi, S. (2015). Théorèmes limites et ordres stochastiques relatifs aux lois et processus stables (Doctoral dissertation, Université Toulouse 3 Paul Sabatier).
- [104] R. Mikulevičius and F. Xu. On the rate of convergence of strong Euler approximation for SDEs driven by Levy processes. *Stochastics*. 569-604 (4), 2018.

- [105] N. Fournier. On pathwise uniqueness for stochastic differential equations driven by stable Lévy processes. *Annales de l'IHP Probabilités et statistiques*. 138-159 (49), 2013.
- [106] Ripley, B. D. (2005). *Spatial statistics*. John Wiley and Sons.
- [107] Kuruoglu, E. E. (2001). Density parameter estimation of skewed stable distributions. *IEEE Transactions on signal processing*, 49(10), 2192-2201.
- [108] Samoradnitsky, G. (2017). *Stable non-Gaussian random processes : stochastic models with infinite variance*. Routledge.