Habilitation à diriger des recherches

Université de Rennes 1

The hidden part of Markovian stochastic processes for biology and ecology

---

presented by
Marie-Pierre Etienne

Jury:

Olivier Gimenez	Directeur de recherche, CNRS, CEFE, France	Examinateur
Mevin Hooten	Professor, Colorado State University, USA	Reviewer
Valérie Monbet	Professeure, Université de Rennes 1, France	Examinatrice
Anne Philippe	Professeure, Université de Nantes, France	Rapporteure
Pierre Pudlo	Professeur, Université Aix Marseille, France	Rapporteur
Stéphane Robin	Professeur, Sorbonne Université, Paris, France	Examinateur

Defended in Rennes, on

October 15, 2021

Contents

List of Figures

List of Tables

Acknowledgments

I would like to begin by thanking Mevin Hooten for agreeing to review my manuscript and to join on my jury. His expertise in statistical ecology and especially in movement ecology makes his opinion very valuable to me, especially since I spent quite some time reading his papers. I will now switch to French for the remaining of this Acknowledgment chapter to feel a bit less constraint by the language.

Je voudrais remercier également Anne Philippe et Pierre Pudlo, qui ont accepté de rapporter cette habilitation bien que la part bayésienne ne soit sans doute pas assez conséquente à leur goût. Je suis ravie de pouvoir bénéficier de votre regard sur mon travail et ainsi de votre compétence en processus stochastiques, en statistique bayésienne et en statistique computationnelle.

Valérie Monbet est directrice de l'équipe de statistique ici à Rennes, impliquée dans de nombreux projets de statistiques qui, pour beaucoup sont liés au milieu marin. Elle est spécialiste des modèles markoviens, souvent plus gourmands que les modèles auxquels je me suis intéressée mais notre proximité dans les outils mathématiques utilisés, dans le domaine d'application ainsi que géographique rend son avis précieux. Merci d'avoir accepté de participer à ce jury et j'espère que nous aurons bientôt l'occasion de travailler ensemble.

Olivier Gimenez est un statisticien écologue ou un écologue statisticien, il a d'ailleurs monté le GdR Ecologie statistique. Il réussit parfaitement à marier statistique, informatique, écologie et enjeux sociétaux et fait de ce mariage à 4 (à la moralité potentiellement discutable) une réussite inspirante. Merci donc de participer à ce jury et de lui donner une petite coloration d'écologie.

Mes années à AgroParisTech m'ont permis de rencontrer des nombreux collègues et j'y reviendrai plus tard et il me tenait à coeur qu'AgroparisTech soit représenté dans mon jury. Manque de chance, Stéphane Robin a depuis décidé de laisser partir AgroParisTech à Saclay sans lui. Je n'ai donc plus de représentant d'AgroParistech mais je suis tout de même ravie, Stéphane que tu ais accepté de présider ce jury et j'attends avec une impatience mélée d'une petite pointe d'angoisse les remarques toujours pertinentes mais parfois destabilisantes que ta curiosité naturelle va amener sur mon travail. Je n'oublie pas que tu m'as toujours encouragée à m'engager dans la rédaction de cette habilitation en m'expliquant à quel point cet exercice était satisfaisant et si je t'avais écouté un peu plus, j'aurais sans doute moins trainé.

J'ai croisé de nombreuses personnes dans mon parcours, qui m'ont aidée, fait grandir et/ou avec qui j'ai collaboré et je souhaite prendre le temps de les remercier puisque le nombre de pages n'est pas limité. J'espère n'oublier personne, mais si jamais, je m'en excuse

d'avance. Pour essayer de mettre un peu d'ordre, j'ai choisi de construire ces remerciements comme une frise chronologique.

Ma vie de recherche a commencé à Nancy, à l'Institut Elie Cartan avec Pierre Vallois qui avait accepté d'encadrer mon travail de thèse. Pierre merci d'avoir contribuer à faire entrer un peu de rigueur dans mon esprit fantasque, merci pour ton accompagnement pendant la thèse et merci de m'avoir aidé à assumer mon gout pour les recherches à l'interface entre les disciplines. Je suis heureuse que nous ayons eu, à nouveau, récemment, l'occasion de travailler ensemble et de se trouver un sujet suffisamment retors pour que cette collaboration se poursuive longtemps.

A la fin de ma thèse, j'ai rencontré une personne formidable, qui m'a largement démontré que les statistique appliquées n'était pas les statistiques du pauvre. Une personne que sa curiosité intellectuelle a conduit à monter un laboratoire à l'interface des statistiques et de la génomique, à entretenir des collaborations régulières avec des médecins, à apprendre le Nahuatl et inspirer de nombreux chercheurs. Bernard Prum m'a vraiment aidée à construire une identité de recherche. Merci beaucoup Bernard de m'avoir accueillie durant les 18 mois que j'ai passés au Laboratoire Statistique et Génome, à une période où j'hésitais beaucoup à poursuivre la recherche dans un cadre académique. L'environnement que tu avais su créer à Evry et, bien sûr, tous les membres du laboratoire à cette époque (notamment Hugues, David, Vincent et Catherine) étaient propices au travail, à l'esprit d'équipe et à la bonne humeur.

Et puis, ensuite ça été l'ENGREF, feue l'Ecole Nationale du Génie Rural, des Eaux et des Forêts, en promettant de ne plus m'occuper d'application en génomique. Avant même mon premier jour dans ce nouveau poste, Eric Parent m'avait proposé de l'accompagner à la soutenance de thèse de son étudiant, il s'agissait d'Etienne Rivot (je reviendrai sur son cas plus tard), thèse soutenue à Agrocampus à Rennes, dans l'amphi même où je soutiens mon Habilitation, coïncidence ???? sans doute oui, mais une drôle de coïncidence. Merci beaucoup, Eric de m'avoir fait partager ton goût des applications notamment en écologie, de m'avoir initiée au raisonnement bayésien, un cadre qui rend le chemin des probabilités vers les statistiques bien moins destabilisant intellectuellement. Tu n'as pas tout à fait réussi à me convertir au Bayésianisme, mais tu m'as donné le goût de la modélisation, des approches hiérarchiques, des aspects computationnels et du questionnement en lien avec l'écologie. Tu m'as permis très vite de participer à un encadrement de thèse. Eric, merci aussi de m'avoir donné le goût des statistiques au sommet et ceci malgré ce mur de la honte qui me donne mal au ventre à peu près une fois tous les deux ans.

A l'ENGREF et avant qu'on déménage à AgroparisTech, j'ai rencontré mon voisin de bureau Gabriel Lang (et souvent aussi Paul Doukhan qui venait trainer dans mon bureau quand il ne te trouvait pas). Je ne me suis pas rendue compte immédiatement à quel point cette rencontre serait marquante et ceci pour des raisons très variées. Merci Gabriel, pour les heures passées dans mon bureau à essayer de m'empêcher de travailler avec tes discussions fantasques sur des sujets on ne peut plus variés, au hasard : la litérature, le vin, le musée des boutons, merci pour les éclats de rire partagés et puis aussi merci pour m'avoir embarqué dans le voyage de longue haleine vers les excursions de Ornstein et Uhlenbeck. Merci de rester tellement imprévisible que tu es aujourd'hui en charge du numérique alors que tu

m'avais si longuement expliqué qu'il était hors de question que tu passes ta vie à appuyer sur des boutons d'ordinateur.

Mon séjour à UBC, dans l'équipe de Murdoch Mc Allister a été extremement important dans ma découverte de l'halieutique et pour me permettre de développer des outils plus proches des décisions opérationnelles. Merci Murdoch de m'avoir accueilli dans ton laboratoire. Cette mobilité d'une année a été motivée par mon souhait de m'immerger de façon plus proche des applications mais deux personnes m'ont poussée à sauter le pas. Tout d'abord Avner Bar Hen, qui lorsque je lui ai fait part de mon envie de partir à l'étranger m'a trouvé dans les deux jours qui suivaient un point de chute possible dans un laboratoire prestigieux. J'ai finalement été ailleurs, mais ce coup de pouce a été décisif et m'a permis de me lancer. Depuis Avner, on se croise souvent, on a même passé pas mal de temps ensemble, parfois dans des conditions rocambolesques, j'espère que nous aurons un jour l'occasion de travailler vraiment ensemble. Je ne sais pas si ca sera très efficace, mais je sais qu'on y prendra un grand plaisir. Une deuxième personne m'a aidée à construire ce projet de départ pour UBC, alors que je n'avais jamais quitté la France. Merci Liliane, pour m'avoir provoquée en me disant, alors que je parlais pour la n$^{\text{ième}}$ fois de mon souhait d'une mobilité à l'étranger, que je serais encore là dans 10 ans. Cette petite piqûre à mon amour propre m'a bien poussée à te prouver le contraire. Finalement, ta prédiction s'est révélée deux fois fausse, une première fois car j'ai bien fait cette mobilité et une deuxième car, avant l'échéance des 10 ans, j'étais partie m'installer à Rennes. Merci pour cet encouragement et merci aussi d'avoir participé à mon recrutement.

Ces remerciements commencent à s'éterniser et pourtant j'en arrive seulement à remercier les collègues avec qui j'ai partagés mon quotidien pendant plus de 10 ans à AgroParisTech. Tout d'abord, Emilie, avec qui j'ai beaucoup travaillé en enseignement, un peu en recherche et avec qui je partage une longue amitié (et quelques verres). Ensemble, nous avons choisi d'accompagner Tristan Mary Huard dans son exil, quand, dans une tentative, un peu desespérée de trouver des bureaux confortables à AgroParisTech, on l'avait envoyé dans la paillotte sur le toit. Je garde un excellent souvenir de cette période, durant laquelle on partageait à 4, Marie Laure faisant partie de la dream team, ce bureau très mal chauffé, perdu dans le département qui s'appelait encore Zootechnie. Plus généralement le laboratoire MIA-Paris regorgent d'enseignants, de chercheurs avec qui il fait incroyablement bon travailler. Pierre Barbillon, le plus Parisien des Lorrains fait preuve d'un flegme sans pareil matiné d'un humour mordant (mais jamais trop fort). Lorsqu'il est accompagné de Julien Chiquet (Code Gourou) et de Sophie Donnet (on relève les manches et on fait les calculs), les vannes volent bas. C'est le moment de garder sa susceptibilité dans sa poche, sa langue pendue et de se préparer à rire beaucoup. Cette joyeuse équipe n'est qu'une petite partie du laboratoire MIA-Paris, un laboratoire qui allie curiosité scientifique et bonne humeur, bref il y fait très bon vivre. De nombreuses personnes participent à ce climat, mais une réussite spectaculaire réside dans les liens très étroits entre les permaments et les doctorants, et ce malgré les renouvellements des générations, notamment Caroline toujours pétillante; Frédéric et ses contrepétries; Anna, vive intelligente et drôle; Loïc qui excelle tout autant en statistique qu'en design; Marie C, toujours le sourire et d'une gentillesse à toute épreuve (au point de supporter même les néo Parisiens); Marie P, dont l'énergie sans égale participe à

augmenter le niveau de bonheur moyen autour d'elle. J'insiste beaucoup sur les qualités humaines de chacun, mais je n'oublie pas que ce laboratoire regorge de gens brillants, curieux et ouverts avec qui le travail avance et donne lieu à de beaux résultats.

J'ai de nombreuses applications dans le domaine de l'halieutique, sans aucun doute car j'aime particulièrement travailler avec Etienne Rivot (rappelez vous, je le connais depuis qu'il est tout petit, quand il avait soutenu sa thèse), Stéphanie Mahévas et Nicolas Bez, cette petite bande d'halieutes qui aime particulièrement la modélisation, presqu'autant que le ski ! On travaille très bien ensemble que ce soit pour encadrer des doctorants, ou pour créer des groupes scientifiques avec des noms qui font notre fierté[1] et puis on y prend plaisir.

En 2017, j'ai quitté AgroParisTech et Paris, pour le climat doux et océanique de la Bretagne. Quitter un tel environnement de travail a constitué une petite source d'inquiétude, j'ai du apprendre à travailler dans un bureau confortable, avec une fenetre qui ferme et je dois dire qu'au début, je n'avais pas grand monde qui faisait irruption dans mon bureau ... un véritable choc. Heureusement toute l'équipe de statistique s'est mobilisée pour m'aider à m'acclimater : ils se sont forcés à venir discuter dans mon bureau, à lancer des vannes. Merci à tous pour l'accueil que vous m'avez réservée. Plus spécifiquement, merci David d'avoir évoqué ce poste devant moi et de ton soutien lors de ma candidature, merci à Mathieu et Jean-Louis pour leur relecture du manuscript[2]. Merci Sébastien pour nos discussions variées et pour tenter de mettre en moi un semblant de début d'esprit de compétition[3] Magalie, merci de ta disponibilité et de ta réactivité non seulement pour nos petits problèmes informatiques mais aussi lors de notre dernière collaboration épique. Merci à Héléna, pour ta présence toujours souriante et pour ton aide dans l'organisation en général et pour ma soutenance. Enfin, un très grand merci François, pour ton opiniatreté à me pousser à rédiger cette habilitation et pour ton aide et tes encouragements tout au long de la rédaction. La (légère) pression plus que bienveillante que tu m'as mise pour que j'avance ce projet a été décisive. Merci.

Je voudrais (presque) conclure en remerciant plus particulièrement les doctorants avec qui j'ai eu l'occasion de travailler dans le cadre d'un encadrement officel ou non. La thèse de Sophie Ancelet a été l'occasion de m'initier à la modélisation bayésienne, aux données zero inflated et à l'encadrement. Travail que j'ai poursuivi avec plaisir avec Jean-Baptiste Lecomte, qui a rejoint aujourd'hui cette petite bande d'halieutes qui aime la modélisation, j'espère qu'on aura l'occasion de travailler bientôt à nouveau ensemble. Le travail de thèse de Maximilien Simon a été pour moi, l'occasion d'une plongée, un peu plus profonde en halieutique, et malgré la distance on a réussi à bien travailler ensemble. La preuve, Max, tu as arrété de faire de la recherche pour mieux la diriger ! Le travail avec Rémi Patin, lorsqu'il était encore en thèse, m'a permis de quitter le monde marin et de travailler sur des déplacements des zèbres. Merci Rémi, car honnêtement illustrer une présentation avec un poisson ou un zèbre, ça ne rencontre pas du tout le même accueil. Tu dois me comprendre maintenant que tu travailles sur le saumon ! J'ai replongé dans le Golfe de Gascogne, au

[1] Moving2Gather, pour un groupe sur les modèles statistiques pour l'écologie du déplacement, c'est bien quand même ! Ceci dit la team AgroParisTech, est à l'origine de FinistR et StateOfTheR, c'est bien aussi.

[2] et pour ton engagement sans faille à venir refaire le monde dans mon bureau.

[3] Mais, non ! Je ne viendrai pas aux cours de Self défense faire des pompes!

milieu des soles, des merlus et mêmes des langoustines pour travailler avec Baptiste sur des modèles intégrant les captures commerciales pour mesurer l'abondance. Baptiste, ton intelligence, ta curiosité, ton soucis du travail bien fait rendent la collaboration avec toi, on ne peut plus agréable, j'ai presqu'envie que ta thèse ne se fasse pas en 3 ans... Enfin je vais finir, ce paragraphe en remerciant Pierre Gloaguen qui a réussi à obtenir une thèse en écologie en travaillant presqu'exclusivement sur des équations différentielles stochastiques. Pierre, c'est extrêmement difficile d'exprimer à quel point j'ai eu plaisir et j'ai toujours plaisir à travailler avec toi. Tu es curieux, passionné, particulièrement malin, pas du tout sarcastique, toujours mesuré et précautionneux, opérationnel dès le matin. Sur un plan plus personnel, j'ai adoré partagé avec toi ta passion pour le cacao, l'eau gazeuse et je continue à profiter de ton grand talent pour la belote. J'espère qu'on va continuer à partager ce genre de projet. Mentionner Pierre, appelle immédiatement à mentionner Sylvain Le Corff, qui se retrouve de fait dans cette petite partie sur les doctorants, ca rajeunit non Sylvain ? Je sais que tu trouves ta place en toute cirsconstance. J'ai été bien heureuse de travailler avec toi, mais aussi sur un plan plus personnel, des bonnes soirées que nous avons pu passées avec Pierre.

Je vais finir cette (très) longue partie en remerciant Les personnes les plus proches de moi. Merci Papa pour participer encore une fois à la préparation du pot qui va suivre et bien sûr pour tes encouragements[4]. Maman, j'aurais aimé que tu sois là et je pense fort à toi. Mon quotidien est animé par des petits (enfin plus tant que ça) êtres vifs et bavards, débordants d'énergie et de questions et qui rendent la vie un peu Rock n Roll mais tellement plus chouette, merci Maïa et Théo pour votre énergie débordante. Enfin, un merci très spécial à Cyril, mon git Gourou, qui m'accompagne et m'encourage depuis un moment maintenant, et qui va gérer l'organisation des festivités à suivre. Il a promis, de lire cette habilitation mais au moment où j'écris ces lignes je crois qu'il est en train de refuser l'obstacle. Cyril, si tu me lis (et sinon tant pis), merci pour ces 15ans et plus passés ensemble, merci pour ces deux beaux enfants et merci pour partager ma vie et mes humeurs (changeantes) au quotidien.

En voyant la longueur de cette section écrite en Français, je me rends compte que le choix d'écrire ce document en anglais a sans doute sauvé quelques arbres. Je vous souhaite une bonne lecture.

[4] *Alors, elle en est où ton HDR ?*

Chapter 1

Introduction

I have enrolled for a PhD in Probability in 1999[1], with a mathematical background but the ambition to work at the interface between Mathematics and Biology. My initial interests focused on the development of methods for biological sequence analysis and then turned to the development of methods for movement ecology or population monitoring.

I define myself as a researcher in statistics working at the interface with biology. I find it particularly exciting to formalize a biological problem into a probabilistic or statistical problem, build up some mathematical answers and then being able to come back to biologists and propose some concrete solutions (which usually requires a few trips back and forth). Obviously, the proposed formalization and the probabilistic tools involved are deeply influenced by my research background and in progressing through this document it will become evident that continuous time Markovian processes play a predominant role in my approach to the biological questions of interest.

I would like to mention that the research activities, I will describe in this document, are the results of my interests for sure, of some opportunities certainly, and also, for an important part, the result of my interactions with amazing colleagues (with whom I have worked in laboratories that deserve the label 'Great place to work'). I started my research journey in a mathematical lab, under the supervision of Pierre Vallois who taught me how important it is to propose a clear mathematical formulation of a problem and how the fine properties of stochastic processes are important to solve applied questions. I had then the chance to spend 18 months in the 'Statistics and Genome' lab and work with Bernard Prum who was an amazing pedagogue and researcher. He was not only an excellent statistician, but also had an extensive comprehension of biological mechanisms. He had a talent to interact with researchers from other disciplines, and also with students. Thanks to him, I have understood that there was no opposition between applied and theoretical statistics but some sort of continuum and we were free to choose different positions on this continuum depending on our objective and personal interests. He also showed me how fun it is to teach mathematics to students from different fields and how to help them to overcome their fear

[1] As I started the writing of this document, I went back to my PhD thesis and read the introduction again : the third sentence stated that the human genome was about to be achieved! I suddenly realized that I should hurry up if I expect to finish this 'Habilitation à diriger des recherches' before the human genome is fully understood. Fortunately it takes longer to understand the genome than to read it.

of mathematical formulas by illustrating them with careful chosen examples. I have spent more than 10 years in the MIA Paris lab where I have found a very inspiring and inclusive environment. This lab is full of talented and inspiring scientists, at different position of the continuum I mentioned before. But more than this collection of smart people, this lab is the place where you enjoy to come and where the collaborations between people are natural. I have written papers with almost half of the people of the labs and I'm still collaborating with them and I have always a tremendous pleasure to visit this lab. I arrived 4 years ago in the statistical team of IRMAR in Rennes and again I have been lucky enough to meet very nice colleagues, from which I'm starting to learn statistical learning but also the power of exploratory multivariate data analysis. These colleagues have been supportive enough to convince me to finally write this Habilitation.

The purpose of this last paragraph was to highlight that my research activities has benefit of many different influences that I have tried to process but as a (soon to be) famous researcher remarked to me[2], 'Les chiens ne font pas des chats'[3] and the researcher I have become is (at least partially) the product of these different influences.

1.1 How can stochastic modeling contribute to biology?

As a young PhD student (and still is today), I was convinced that my research should have direct applications. The choice of applications in biology was initially a question of opportunity, and I could also have worked at the interface between mathematics and finance[4]. But it turned out that I found biology fascinating and I realized the need for methods to help analyze the complex phenomena that biologists were trying to understand. The Oxford dictionary defines Biology as *The study of living organisms, divided into many specialized fields that cover their morphology, physiology, anatomy, behavior, origin, and distribution.* I have focused mainly on two aspects of this field: genomics and ecology.

The denomination genomics appears in 1986 [Yad07] and stands for the science which studies the structure, function and evolution of genomes and proteins. This field has become more and more prominent in the last decades and has been at the origin of major advances in medicine, vegetal and animal production. This progress has been made possible thanks to research at the interface between different disciplines such as biology, physics, computer science and mathematics. As part of the methodological development in genomics, we found the works around the detection of patterns in biological sequences. This specific issue in the field of genomics has been examined using computational and algorithmic methods and/or probabilistic methods: BLAST software, as an example, is a powerful tool for sequence alignment and uses a stochastic model of sequences to evaluate the statistical significance of local alignment scores based on Karlin, Dembo, and Kawabata [KDK90] work.

[2] Pierre, I mean you!
[3] Literally 'Dogs don't make cats' or more accurately 'Apple does not fall far from the tree'.
[4] Actually no, I'm kidding!

Ecology is the branch of biology which studies the relationships between organisms and their environment. The increase in the world's human population and urbanization are leading to major changes in land use and land cover [Gri+08]. As a consequence the threats on biodiversity are also growing [Höl+10; WK10; Gal+14]. The conservation of biodiversity and the sustainable management of exploited populations are therefore major societal challenges. In this document, I focus on two specific aspects of conservation at the population status assessment stage: understanding the use of space by individuals and monitoring the abundance of a population. Movement and abundance evolution involve complex systems that are :

- multi factorial,
- highly variable,
- dynamic,
- and might not even be directly observable.

Stochastic modeling is an interesting entry to progress in the understanding of these systems and/or help in the prediction of their evolution.

1.1.1 How to propose relevant models?

Proposing and studying a model for a new biological question is always a trade-off between two contradictory desires:

- the desire to propose realistic model,
- the desire to propose tractable model.

The compromise between these two conflicting desires must be driven by the end use of the model:

- Going for a simple model allows to explore its mathematical properties from a theoretical prospective. Beyond the mathematical interest of such an exploration, it is useful to propose the correct statistical test or an efficient computational algorithm. A good illustration is undoubtedly the Wright-Fisher's model and Kingman's coalescent [Kin82]. Despite its apparent simplicity it has given rise to a great deal of research in probability ([Tav84; TV+09; Fu06] among many others) but is also widely used to test different hypotheses concerning heredity mechanisms in different populations [DT95; Fu96].
- When biological knowledge has already identified important drivers for different mechanisms, the proposed models must account for them. This often leads to more complex models for which the issue is not so much the study of the mathematical properties but rather to propose efficient, preferentially unbiased, estimation methods that can be used in practice.

The real life of an applied statistician working with biologists and ecologists is some kind of nice but mostly unpredictable journey between models with different level of complexity depending on the question of interest.

1.1.2 Statistical or mechanistic model?

For a long time, there has been a dichotomy between statistical models which are more adequately named phenomenological models in opposition to mechanistic models. Following [Bol08], the first modeling framework concentrates on observed patterns in the data, while the second one focus on the underlying processes which are responsible for the pattern in the data. Mechanistic models were argued to exhibit better predictive performance especially in a context of a changing environment [Eva12]. However, some mechanistic models tend to represent any single process and might lead to incredibly large models. As a consequence, the estimation of the many parameters involved in such models is challenging, if not impossible (due to the unidentifiability of the model, the lack of informative data or both), and the benefit of the approach is lost. However, the strict boundary between the statistical model and the mechanistic model that existed tends to be blurred and the two approaches can be reconciled within models integrating different components.

The main purpose of the models I have considered so far has been to understand (and more rarely to predict), potentially unobserved, biological processes. These models usually have a mechanistic as well as a statistical component, and I will use the generic term stochastic models throughout this document.

1.1.3 The connection with ecologists

In his Habilitation thesis, my colleague Mathieu Emily [Emi16] mentioned the *forward-backward algorithm of biostatistics*. Working with genome wide association studies, he meant that new technological developments were opening up new biological questions; this new type of data gave rise to new needs in terms of statistical methods. Although technological change is somewhat slower in ecology[5] than in genomics, I will borrow his terminology to emphasize the needs of close interactions between statisticians and their field of applications. Biological questions will feed into new research questions in statistics (more generally in mathematics), as progress in mathematics can itself open up new questions in biology. This close relationship

- enables statisticians to understand the biological issues and the data collected and thus to propose suitable models for which only biologically reasonable (as much as possible) simplifications are proposed,

- is an opportunity to suggest new issues for which the statistician has identified relevant approaches or tools,

[5]Even if movement ecology is experiencing fast technical development.

- sometimes takes the statistician into the field to assist[6] in the data collection and assures him or her to remain humble.

1.2 The key ingredients

Proposing suitable methods for analyzing biological data requires an understanding of the issues but it also requires the use of relevant probabilistic tools and methods. Most of the problems I have studied so far, have a temporal and/or a spatial component, so naturally I have focused on the study of stochastic processes. My initial training in probability and, undoubtedly, the great influence of my PhD advisor, Pierre Vallois, allowed me to appreciate the richness of Markovian models. These models, both flexible and with well known mathematical properties, are not only relevant in many applications but also offer the potential for theoretical or methodological results and thus the development of efficient estimation methods.

These models lack to accommodate abrupt changes in dynamics. The use of hidden component in the modeling approach is one possible approach to address the issues of regime switching and more generally, hidden (also called latent) layer is the cornerstone of Hierarchical modeling which is now a standard in statistical ecology. Hidden (potentially Markovian) variable models are a major research topic of the MIA Paris team in which I spent more than 10 year.

Stochastic Markovian processes and hierarchical model are the two key ingredients of my research.

As I will detail later, this document is organized by field of applications, however, most of the stochastic processes are common to the different application fields and I decided to include their presentation in the introduction.

1.2.1 Markovian Stochastic processes

Spatial organization along the genome induces potential dependence just like the dependence one would expect to observe when studying displacement.

This section introduces the different Markov processes that will be used in the next chapter. A very nice and complete introduction to these processes can be found in Karlin [Kar14] and Karlin and Taylor [KT81].

A homogeneous **Markov chain** on a finite state space $\mathscr{S}$, $\boldsymbol{X} := (X_n, n \in \mathbb{N})$, is a discrete time model which verifies for all $(i,j) \in \mathscr{S}^2$ the two following properties:

$$\begin{aligned}\mathbf{P}\left\{X_{n+1} = j | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\right\} &= \mathbf{P}\left\{X_{n+1} = j | X_n = i\right\} \quad \text{(Markov)},\\ &= \mathbf{P}\left\{X_1 = j | X_0 = i\right\}, \quad \text{(Homogeneity)}.\end{aligned}$$

The matrix $P = (p_{ij})_{(i,j)\in\mathscr{S}^2}$, where $p_{ij} = \mathbf{P}\left\{X_{n+1} = j | X_n = i\right\}$ is called the transition matrix.

[6] dare I say participate

The sojourn time in state i, defined as $\min\{n, X_n \neq i | X_0 = i\}$, follows a geometric distribution with parameter p_{ii} and the sequence of sojourn times forms a sequence of independent random variables (rv's). This property might have consequences and will be discussed in the Chapter 3.

As I will attempt to illustrate throughout this document, it is often more relevant from a conceptual and computational point of view to consider continuous-time processes. $\boldsymbol{X} := (X_t, t \geq 0)$ is a **continuous time homogeneous Markov chain** on a finite state space $\mathscr{S}$ if

$$\mathbf{P}\left\{X_{s_{n+1}} = j | X_{s_0} = i_0, \ldots, X_{s_n} = i\right\} = \mathbf{P}\left\{X_{s_{n+1}} = j | X_{s_n} = i\right\} = P_{ij}(s_{n+1} - s_n).$$

Similarly to discrete time Markov chain, a continuous time Markov chain can be defined through its transition function $P(.) = (P_{ij}(.))_{(i,j)\in\mathscr{S}^2}$.

$$P_{ij}(s) = \mathbf{P}\left\{X_s = j | X_0 = i\right\}.$$

Equivalently, as found in Theorem 2.8.2 in Norris [Nor98], it can be defined through its infinitesimal generator Q:

$$Q = \lim_{h\to 0^+} \frac{P(h) - I}{h} = \begin{pmatrix} -q_1 & q_{12} & q_{13} & \cdots & q_{0S} \\ q_{21} & -q_2 & q_{23} & \cdots & q_{2S} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & q_{S-1\,S} \\ q_{S1} & \cdots & \cdots & q_{S\,S-1} & -q_S \end{pmatrix}, \qquad \text{(Eq 1.2.1)}$$

with $q_{ij} \leq 0$ for all $(i,j) \in \mathscr{S}^2$ and $q_i = \sum_{j\in\mathscr{S}, j\neq i} q_{ij}$. As for the discrete time model, the sequence of sojourn times is a sequence of independent rv's. The sojourn time in state i follows an exponential distribution with mean $1/q_i$. In Chapter 2, we will consider a continuous time Markov chain to model the alterations in a cohort of patients.

A large part of my work makes use of continuous time processes defined on a continuous state space. **Brownian motion** (BM) is to stochastic processes what normal variables are to random variables. There exist many definitions or constructions of the Brownian motion. Following Revuz and Yor [RY13], a one-dimensional standard Brownian motion is defined as *an almost surely continuous process* $\boldsymbol{W} = (W(t), t \geq 0)$ *with* $W(0) = 0$ *and independent increments such that for each* t*, the random variable* W_t *is centered, Gaussian and has variance* t*.* A d-dimensional Brownian motion is defined as $d-$collection of independent one dimensional Brownian motion. I will use the same notation $\boldsymbol{W}$ all over this document for both one and d-dimensional Brownian motion. The process $\boldsymbol{W^\mu} := (W(t) + t\mu, t \geq 0)$ is named **Brownian motion with drift**.

When analyzing movement data, it is often simpler to propose a model for the speed than a model for position. Nevertheless, speed can be influenced by the environment in which the individual is located and therefore by his position. We therefore seek for a model linking speed and position, which is the classical situation of differential equations. The stochastic analog of differential equation is Stochastic Differential Equation (SDE). The integration

has to be understood in the Itô's definition (a very gentle and practical introduction of SDE is given by Iacus [Iac09]). A process $\boldsymbol{X}$ is said to be a weak[7] solution to a SDE, if there exists a probability space where $\boldsymbol{X}$ verifies:

$$X(t) = X(0) + \int_0^t b(X(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\,\mathrm{d}W(s).$$

The SDE is mostly presented with the corresponding differential form:

$$\mathrm{d}X(s) = b(X(s))\,\mathrm{d}s + \sigma(X(s))\,\mathrm{d}W(s), \tag{Eq 1.2.2}$$

the initial condition X_0 can be a constant or a rv (independent of the σ-algebra associated to $\boldsymbol{W}$). As the drift and diffusion terms do not depend on time, the solution is homogeneous in time. The existence of a weak solution is guaranteed under some mild regularity conditions for b and σ, and the solution is a Markovian continuous time process. However, the transition density of this process is in general unknown. The question of the estimation of such processes will be examined in section 3.3.3.

The Brownian motion with drift $\boldsymbol{W}^{\boldsymbol{\mu}}$ defined before can also be defined as the solution to the following SDE:

$$\mathrm{d}X(s) = \mu\,\mathrm{d}s + \mathrm{d}W(s), \quad X(0) = 0.$$

As already mentioned, among the class of processes defined to be the solution to a SDE, few have explicit transition densities function but so is the one dimensional **Ornstein Uhlenbeck** (OU) process, $\boldsymbol{U}_{\tau,\sigma}^{\mu,U_0}$, defined as the solution to the following SDE:

$$\mathrm{d}X(s) = -\tau\,(X(s) - \mu)\,\mathrm{d}s + \sigma\,\mathrm{d}W(s), \quad X(0) = U_0, \tau > 0. \tag{Eq 1.2.3}$$

It can also be defined as the unique continuous Gaussian process whose covariance function verifies:

$$Cov\left(U_{\tau,\sigma}^{\mu,U_0}(t), U_{\tau,\sigma}^{\mu,U_0}(s)\right) = \frac{\sigma^2}{2\tau} e^{-\tau(t-s)}\left(1 - e^{-2\tau s}\right), \quad 0 \leq s < t.$$

The transition density function of this homogeneous continuous time continuous space Markov model, is a known normal density function:

$$U_{\tau,\sigma}^{\mu,u_0}(t) \sim \mathcal{N}\left(\mu + e^{-\tau t}(u_0 - \mu), (2\tau)^{-1}\sigma^2(1 - e^{-2\tau t})\right),$$

and as a consequence the simulation and the estimation of this process is straightforward.

This process also admits a unique normal stationary distribution centered in μ with variance $(2\tau)^{-1}\sigma^2$. In Chapter 2, we focus on a specific unit variance stationary OU process, $\boldsymbol{U}_\tau$ solution to:

$$\mathrm{d}U_\tau(s) = -\tau U_\tau(s)\,\mathrm{d}s + \sqrt{2\tau}\,\mathrm{d}W(s), \quad U_\tau(0) \sim \mathcal{N}(0,1), \tau > 0. \tag{Eq 1.2.4}$$

To conclude, this collection of stochastic processes open a wide variety of modeling possibilities, however individual movement tend to present non stationnarity and it can not reasonably be modeled as a single Markov process.

[7] A more formal definition of stochastic differential equations and the difference between strong and weak solutions can be found in Karatzas and Shreve [KS98].

1.2.2 Hidden variables

Biology and ecology involve systems with complex structures whose responses to different drivers occur at various scales of time and space. Most often, these systems are only partially observed and the data resulting from these observations exhibit complex dependency structures. One of the challenges of stochastic modeling lies in developing parsimonious models that account for these dependencies. The use of hidden variables makes it possible to represent these dependencies by combining simple elements that are only partially observed. The simplest example of hidden variable model is probably the linear mixed model. This model has been widely used, by instance, in genetics where linear mixed models are used to quantify the variation of a phenotypic trait in the population and also to account for kinship relationship among individuals. A classical formulation of a mixed effect model is given by

$$Y = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{Z}X^{8} + E, \quad X \sim \mathcal{N}_p(0, \sigma_X^2 I_d), \quad E \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

where

- Y is a vector of dimension n and stands for the phenotypic trait of interest measured on n individuals,
- $\boldsymbol{H}$ is the design matrix of the fixed effects, those effects being captured by $\boldsymbol{\theta}$,
- $\boldsymbol{Z}$ is the design matrix of the random effect (indicating the linkage relationship by example), captured by the Gaussian vector X (X being understood in this context, as the unknown traits inherited from the relatives).

Conditionally on X, the vector $(Y_i)_{i=1,\dots,n}$ are independent variables but as X is not observed, the structure of dependence for the vector Y is more complex. In this example, two independent Gaussian vectors are sufficient to represent complex dependency in the observation. The dependency between these n observations is explained by the dependency induced by $\boldsymbol{Z}$ and the unobserved variable X.

Models involving hidden variable, often referred to as Hierarchical models, are quite common in biology and have been growing more and more popular in ecology. The terminology of hierarchical modeling has been introduced by Berger [Ber85] and stands for the technique consisting in the decomposition of high-dimension problems into a series of smaller models linked through a collection of hidden variables. Hierarchical modeling has become increasingly important in the last 20 years, especially in the domain of environmental science and will be detailed more precisely in Chapter 4.

The estimation of such models raises some specific issues. In general, the model is built so that the joint probability distribution $p_\theta\{\boldsymbol{X}, \boldsymbol{Y}\}$ of the hidden $\boldsymbol{X}$ and observed variables $\boldsymbol{Y}$ is easily defined. The likelihood could be derived by integrating over the hidden variables, it is however prohibitively time-consuming with a brute-force approach. The estimation of

[8] In Chapter 3, X will be used to denote the hidden variables, therefore, I have chosen to fix this notation from the beginning even if it is quite unorthodox in the context of linear model.

such models is mostly addressed either using the Expectation Maximization (EM) algorithm introduced by [DLR77], either in a Bayesian framework. The EM algorithm consists in optimizing iterately the expected value of the complete-data log-likelihood $\ln p_\theta \{\boldsymbol{X}, \boldsymbol{Y}\}$ with respect to the unknown hidden variable $\boldsymbol{X}$, given the observed data $\boldsymbol{Y}$, that is, at iteration i, the aim is to maximize the function Q defined as :

$$Q(\theta, \theta^{(i-1)}) = \mathbf{E}_{\theta^{(i-1)}} \{\ln p_\theta \{\boldsymbol{X}, \boldsymbol{Y}\} | \boldsymbol{Y}\}.$$

This algorithm converges to a local maximum of the log likelihood. In some specific case, the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$ has an explicit form (as in the random effect model for example), but it is also quite common to use Monte Carlo methods to estimate Q which implies being able to develop efficient algorithm to sample from $\boldsymbol{X}$ given $\boldsymbol{Y}$. This aspect is developed in Chapter 3 in the context of SDE observed with noise. Bayesian framework considers parameters as random variables and the estimation intends to characterize the target distribution, i.e. the distribution of the parameters given the observation $\boldsymbol{Y}$. Except in very simple models, this distribution is not available analytically and simulation approach are used to produce samples from the hidden variables and the parameters as well. Although there exist non iterative algorithm (Importance sampling see Robert [Rob05] by instance), most algorithms represent the target distribution as the stationary distribution of a Markov model and thus propose an iterative simulation of the target distribution. It is quite common to update the set of parameters given the hidden variables and then update the hidden variables given the parameters, especially in dynamical models. Thus the problem of sampling from the hidden states given the parameters and the observed variables is present in both a frequentist and a Bayesian approach. In the specific case, where the hidden variables belong to the Markov Gaussian processes family, [RMC09] have proposed the Integrated Nested Laplace Approximation, a very efficicient method to replace the prohibitive integration by a very good approximation. This method is now widely used as Gaussian Markov fields are flexible enough to account for different sort of spatial dependence and or temporal dependance.

However Gaussian Markov processes are not suitable to model abrupt changes in time series and Hidden Markov Models (HMM) are classically used to addressed this sort of heterogeneity in time [ZML17]. HMM are a two layers models with $\boldsymbol{X}$ an unobserved discrete space Markov chain whose state rules the distribution of the observed second layer $\boldsymbol{Y}$. This class of model is used to propose switching model as illustrated for a movement model in Figure 3.9. HMM are one specific example of hierarchical modeling approach. Hierarchical models raise some specific estimation issues as the likelihood has to be integrated over all possible hidden state, which is often prohibitively time-consuming with a brute-force algorithm. In the case of the HMM, the likelihood is efficiently computed thanks to the forward backward algorithm. The likelihood can therefore be numerically optimized, nevertheless the optimization of the likelihood is rather carried out by the Expectation Maximization (EM) algorithm [DLR77] or one of its variation. This iterative algorithm proposes a method to increase the likelihood without the need to calculate it. Its implementation implies the calculation of expectation on the laws of the hidden variables conditionally to the observed variables which are named smoothing distribution in the context of HMM. Those aspects are developed in Chapter 3.

1.3 Structuration of the thesis

A presentation organized by field of applications or by statistical objects has been the object of a long dilmena. However, since the applications have often been at the origin of my research, I have chosen to preserve this way of thinking about my work right up to the presentation. This document is therefore structured into three main parts.

The first part, presented in Chapter 2, describes my contributions to the detection of atypical portions in the genome. These questions were at the origin of my PhD thesis, during which I worked under the supervision of Pierre Vallois and in collaboration with Jean-Jacques Daudin, who suggested the original problem. From a practical point of view, I have been interested in quantifying to what extent a portion of a signal (a DNA segment for example) is atypical. The **first acceptance of the word atypical** can be introduced as follows. When considering a single sequence, each element of the sequence of interest (DNA, RNA or proteins for example) is assigned a score, defined according to biological knowledge. A scored sequence can then be modeled as a sequence of random variables whose values lie within a finite subset of $\mathbb{R}$. The score of a sequence being defined as the sum of the scores of the elements that make it up, we are interested in the distribution of the local score: the maximum score reached over all the possible subsegments. A segment is considered atypical if its local score is significantly high. The exact distribution of this test statistic being computationally demanding [MD01], we proposed to study its asymptotic distribution under different null models. Genomics is a highly technical field of research, and new technology opens new research challenges. The Comparative genomic hybridization (CGH) profiles measures the quantity of genetic materials along the genome Michels et al. [Mic+07] proposed a review of methods using comparative genomic hybridization (CGH) profiles to understand oncogenesis in a variety of cancer. An alteration of a CHG profile, is a subsegment which exhibits more (or less) genetic material than expected. Detecting this sort of segment shared by a cohort of patients which suffer from the same disease, is a way for the understanding of disease mechanism. This leads to a **second acceptance of the word atypical**. A subsegment will be considered as atypical, if it is long enough and altered in a large proportion of patients. I started to work on this question as the result of a coffee break discussions with Stéphane Robin and Gabriel Lang, two colleagues from MIA Paris. I immediately offered Pierre Vallois to join the project as I knew for sure that he will enjoy this subject. The funny aspect is that the exact same problem was also given attention by Laurent Decreusefond in a context of telecommunication network, at the exact same moment and we are now working together on this question. I have examined the question of the atypicality when the unique sequence of interest is large to characterize the asymptotic distribution and, thanks to the Donsker theorem, this consists in a fine study of Brownian paths properties. When considering a cohort of patients, we have proved that the process of interest converges to an Ornstein Uhlenbeck when the size of the cohort increases. We also proved that the event of interest can be reformulated in terms of the length of the longest excursion above a given threshold. Finally thanks to the study of the Ornstein Uhlenbeck we propose a Monte Carlo methods to compute the probability of this event.

The second part of this thesis focuses on the use of stochastic processes for movement

ecology, the field of ecology that studies the link between animals movement and environment. Movement ecology has experienced a shift in the two last decades thanks to affordable Global Positioning System (GPS) device and their miniaturization. From the first radio collars on grizzly bears from Yellowstone National Park in the 1960s [CC72], the size of the datasets has tended to grow rapidly (a recent study on Adult homing Pigeons, has followed 176 pigeons, corresponding to a total of 8 millions of relocations, [Sch+18]). The link between movement, individual internal states and environment has been formalized by Nathan et al. [Nat+08]. Therefore three of the major questions currently addressed with telemetry data are a) how individuals use space, b) can we infer internal states from tracked movement c) how the use of space is linked to the internal states? These questions were the occasion for various collaborations with colleagues from AgroparisTech (Julien Chiquet, Maud Delattre, Sophie Donnet, Emilie Lebarbier), Agrocampus Ouest (Etienne Rivot), from IFREMER (Stéphanie Mahévas) or from the Centre d'écologie Fonctionnelle et Evolutive (Simon Benhamou) but they were also an opportunity to collaborate with energetic PhD students (Rémi Patin) and even a memorable PhD thesis co-supervision (Pierre Gloaguen). With those great partners, I have developed and I'm developing methodological work to provide analytical tools to ecologist to answer these questions. I have proposed different stochastic dynamical models to represent the movement, including hidden variables to link movement with internal states. In my most recent work, I have been focused on continuous-time stochastic models such as stochastic differential equations. The choice between continuous and discrete models is a matter of debate [McC+14]. I am convinced that despite their apparent complexity, continuous time models provide an interesting approach not only because it raises interesting statistical questions (estimation of, potentially partially, observed SDE, smoothing algorithm, pseudo likelihood estimation of SDE) but also because it allows to differentiate the movement model from the sampling process and, as such, to combine data with different sampling strategy.

The final chapter brings together my works on models for abundance monitoring. In contrast to data obtained from controlled experiments, abundance monitoring data often suffer from problematic characteristics from a statistical point of view that might be adequately addressed by using Bayesian approaches and hierarchical modeling [Cla05]. Thanks to Eric Parent, and in collaboration with Liliane Bel from AgroParistech, Etienne Rivot from Agrocampus Ouest and Hugues Benoît from Fisheries and Oceans Canada, I have been involved in the supervising team of two PhD students, Sophie Ancelet and Jean-Baptiste Lecomte. These thesis were the opportunity to propose models that accommodate zero inflated data with complex dependence structure mainly due to spatial structuration. Commercial fisheries data, not only exhibit the previously mentioned characteristics but also suffer of non random sampling design. I am now co supervising Baptiste Alglave PhD thesis who focus on developing models coupling commercial and scientific fisheries data, taking preferential sampling.

I would like to finish this introduction of my thesis by mentioning that at many occasions, I have collaborated with biologists and ecologists for occasional help in statistics or expertise. Those opportunities to contribute to ecological or biological developments have directly or indirectly contributed to my statistical developments. Likewise, the expert appraisals or

participation in stock assessments have been an opportunity to discover the final aspects of my research and to contribute to its social value. Although these aspects are not developed in this document (they only appear in my publication list), they have been central in defining my identity as a researcher in applied statistics for biology and it is important for me to mention them.

Chapter bibliography

[Ber85] James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 1985.

[Bol08] Benjamin M Bolker. *Ecological models and data in R*. Princeton University Press, 2008.

[CC72] Frank C Craighead and John J Craighead. "Grizzly bear prehibernation and denning activities as determined by radiotracking". In: *Wildlife Monographs* 32 (1972), pp. 3–35.

[Cla05] James S Clark. "Why environmental scientists are becoming Bayesians". In: *Ecology Letters* 8.1 (2005), pp. 2–14.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[DT95] Peter Donnelly and Simon Tavare. "Coalescents and genealogical structure under neutrality". In: *Annual review of genetics* 29.1 (1995), pp. 401–421.

[Emi16] Mathieu Emily. "Contributions to biostatistics: categorical data analysis, data modeling and statistical inference". Habilitation à diriger des recherches. Université de Rennes 1, Nov. 2016. URL: https://hal.archives-ouvertes.fr/tel-01439264.

[Eva12] Matthew R Evans. "Modelling ecological systems in a changing world". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1586 (2012), pp. 181–190.

[Fu06] Yun-Xin Fu. "Exact coalescent for the Wright–Fisher model". In: *Theoretical population biology* 69.4 (2006), pp. 385–394.

[Fu96] Yun-xin Fu. "New statistical tests of neutrality for DNA samples from a population". In: *Genetics* 143.1 (1996), pp. 557–570.

[Gal+14] Alessandro Galli, Mathis Wackernagel, Katsunori Iha, and Elias Lazarus. "Ecological footprint: Implications for biodiversity". In: *Biological Conservation* 173 (2014), pp. 121–132.

[Gri+08] Nancy B Grimm, Stanley H Faeth, Nancy E Golubiewski, Charles L Redman, Jianguo Wu, Xuemei Bai, and John M Briggs. "Global change and the ecology of cities". In: *Science* 319.5864 (2008), pp. 756–760.

[Höl+10] Franz Hölker, Christian Wolter, Elizabeth K Perkin, Klement Tockner, et al. "Light pollution as a biodiversity threat." In: *Trends in Ecology & Evolution* 25.12 (2010), pp. 681–682.

[Iac09] Stefano M Iacus. *Simulation and inference for stochastic differential equations: with R examples*. Springer Science & Business Media, 2009.

[Kar14] Samuel Karlin. *A first course in stochastic processes*. Academic Press, 2014.

[KDK90] Samuel Karlin, Amir Dembo, and Tsutomu Kawabata. "Statistical composition of high-scoring segments from molecular sequences". In: *The Annals of Statistics* 18.2 (1990), pp. 571–581.

[Kin82] John Frank Charles Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13.3 (1982), pp. 235–248.

[KS98] Ioannis Karatzas and Steven E Shreve. "Brownian motion". In: *Brownian Motion and Stochastic Calculus*. Springer, 1998, pp. 47–127.

[KT81] Samuel Karlin and Howard E Taylor. *A Second Course in Stochastic Processes*. Elsevier, 1981.

[McC+14] Brett T McClintock, Devin S Johnson, Mevin B Hooten, Jay M Ver Hoef, and Juan M Morales. "When to be discrete: the importance of time formulation in understanding animal movement". In: *Movement Ecology* 2.1 (2014), p. 21.

[MD01] Sabine Mercier and Jean-Jacques Daudin. "Exact distribution for the local score of one iid random sequence". In: *Journal of Computational Biology* 8.4 (2001), pp. 373–380.

[Mic+07] Evi Michels, Katleen De Preter, Nadine Van Roy, and Frank Speleman. "Detection of DNA copy number alterations in cancer by array comparative genomic hybridization". In: *Genetics in Medicine* 9.9 (2007), pp. 574–584.

[Nat+08] Ran Nathan, Wayne M Getz, Eloy Revilla, Marcel Holyoak, Ronen Kadmon, David Saltz, and Peter E Smouse. "A movement ecology paradigm for unifying organismal movement research". In: *Proceedings of the National Academy of Sciences* 105.49 (2008), pp. 19052–19059.

[Nor98] James R Norris. *Markov Chains*. 2. Cambridge university press, 1998.

[RMC09] Håvard Rue, Sara Martino, and Nicolas Chopin. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series b (statistical methodology)* 71.2 (2009), pp. 319–392.

[Rob05] Christian Robert. *Le choix Bayésien: Principes et pratique*. Springer Science & Business Media, 2005.

[RY13] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Vol. 293. Springer Science & Business Media, 2013.

[Sch+18] Ingo Schiffner, Patrick Fuhrmann, Juliane Reimann, and Roswitha Wiltschko. "Behavioural traits of individual homing pigeons, Columba livia f. domestica, in their homing flights". In: *PloS one* 13.9 (2018).

[Tav84] Simon Tavaré. "Line-of-descent and genealogical processes, and their applications in population genetics models". In: *Theoretical population biology* 26.2 (1984), pp. 119–164.

[TV+09] Jesse Taylor, Amandine Véber, et al. "Coalescent processes in subdivided populations subject to recurrent mass extinctions". In: *Electronic Journal of Probability* 14 (2009), pp. 242–288.

[WK10] David S Wilcove and Lian Pin Koh. "Addressing the threats to biodiversity from oil-palm agriculture". In: *Biodiversity and conservation* 19.4 (2010), pp. 999–1007.

[Yad07] Satya P Yadav. "The wholeness in suffix-omics, -omes, and the word om". In: *Journal of Biomolecular Techniques: JBT* 18.5 (2007), p. 277.

[ZML17] Walter Zucchini, Iain L MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction using R*. CRC press, 2017.

Chapter 2

Stochastic processes and the detection of abnormal regions in biological sequences

Contents

Genomics and Proteomics are interdisciplinary fields of biology focusing on the study of the structure, function and evolution of genomes and proteins. One entry of those sciences is the study of the DNA, RNA sequences or proteins to assess similarity between sequences, predict the 3D structures of proteins, detecting atypical portions of genomes, etc ... This chapter is devoted to the presentation of two different probabilistic methods to detect atypical segments in biological sequences.

The definition of an *atypical* segment differs according to the context. Although the primary motivation arises from biology, the question of detecting atypical segments in sequences concerns many different fields such as telecommunication [FM06] or financial markets. I first

focused on the detection of atypical segments within a sequence, atypicity being measured thanks to a scoring scheme. This question has led to the study of the local score, a process defined to model the maximal score, maximum being taken over all subsegments (Equation Eq 2.1.2). The distribution of the local score under some null model is a corner stone in the construction of a statistical test[1].

Recently, I have been working on the question of *atypical* segment again, but with a different perspective. Considering the genomic profiles of a cohort of patients, a segment will be considered as *atypical* if this portion of genome is altered (i.e presents an excess or a loss of genomic material) in a large proportion of patients.

These two questions are studied in an asymptotic framework: long sequence or large cohort depending on the context and they are both addressed through the study of the properties of two continuous time Markov processes: the Brownian motion and the Ornstein Uhlenbeck process.

The first section of this chapter presents the main results obtained on the local score while second section presents ongoing work devoted to the analysis of atypical segments shared among several sequences.

2.1 The local score for detecting atypical behavior within a sequence

In the genomic context, a sequence is modeled as a sequence of random variables taking values in a finite alphabet $\mathscr{A}$, for example $\mathscr{A} = \{A, T, C, G\}$ when studying DNA sequence, $\{0, 1\}$ in the context of SNPs, or $\mathscr{A}$ the list of the twenty amino acids. This sequence is equipped with a scoring scheme reflecting such desirable properties. An example of a scoring scheme is given in Table 2.1, where each amino acid is associated to a numerical value reflecting its Hydrophobicity (see [KDK90] for different examples); a scoring scheme is just a mapping from $\mathscr{A}$ to $\mathbb{R}$. The sequence of interest $\boldsymbol{X} = (X_k, k \geq 1)$ is the result of this mapping and, as so, is a sequence of real valued random variables, the index k standing for the position in the sequence. In other context X_k could be the load of a network, or the price of an asset.

2.1.1 Definition of the local score

The score of a segment starting at position i up to position j is defined by

$$S_{i:j} := \sum_{l=i}^{j} X_l, \quad S_0 = 0. \tag{Eq 2.1.1}$$

For simplicity, the score $S_{0:i}$ will be simply denoted by S_i.

[1] Although the local score is also used in the context of sequence alignment (between two sequences as in [Wat95], or matching a sequence against a database in [Alt+90]), I won't address this question in this thesis

The local score is the maximum of the scores over all possible segments and is defined by

$$H_n := \max_{1\le i\le j\le n} S_{i:j}, \tag{Eq 2.1.2}$$

The local score terminology is dedicated to the applications in genomics, and few mention can be found in other domain of applications. The local score can also be defined as the maximum of a Lindley process, as in [MD01; DM99]:

$$H_n = \max_{1\le i\le j\le n} (S_j - S_{i-1}) = \max_{1\le j\le n} \left(S_j - \min_{i\le j} S_{i-1} \right) = \max_{1\le j\le n} \tilde{S}_j, \tag{Eq 2.1.3}$$

where $\tilde{S}_j := (S_j - \min_{1\le i\le n} S_{i-1})$ is the Lindley process.

In the context of queuing theory, the terminology of the maximum of a Lindley process is preferred and it represents the longest waiting time experienced by the n^{th} customers arrived in the queue [Igl+72].

Illustration with the human Hemoglobin subunit zeta Proteins are made up of an assembly of 20 base amino acids as illustrated in Table 2.1 for the human Hemoglobin subunit zeta. Different scales are available to measure the hydropathic character of each amino acid, by instance [KD82] proposes the scale presented in Table 2.2. The hydropathic character of a protein is important to understand its function. In particular Intercellular membrane proteins (IMP) represent a class of proteins located in the lipid bilayer of a cell membrane. Those proteins are characterized by the existence of one or several hydrophobic segments. The local score represents the score of the most hydrophobic segment. The distribution of the local score under some null model allows to build a statistical test to assess the significance of the hydrophobic character of a subsegment.

10	20	30	40	50
MSLTKTERTI	IVSMWAKIST	QADTIGTETL	ERLFLSHPQT	KTYFPHFDLH
60	70	80	90	100
PGSAQLRAHG	SKVVAAVGDA	VKSIDDIGGA	LSKLSELHAY	ILRVDPVNFK
110	120	130	140	150
LLSHCLLVTL	AARFPADFTA	EAHAAWDKFL	SVVSSVLTEK	YR

Table 2.1: The amino acid sequence of the human Hemoglobin subunit zeta as found in https://www.uniprot.org/uniprot/P02008

Figure 2.1 presents the different key processes involved in the definition of the local score illustrated with their realizations for the Hemoglobin subunit zeta and the hydrophicity scoring scheme.

First consider the partial sums $\boldsymbol{S}$ and the corresponding running minimum $(\min_{i\le n} S_n)$ which is, by definition, a non increasing process. The corresponding (Lindley) process $\tilde{\boldsymbol{S}}$

R	K	D	E	N	Q	H	P	Y	W
-4.5	-3.9	-3.5	-3.5	-3.5	-3.5	-3.2	-1.6	-1.3	-0.9
S	T	G	A	M	C	F	L	V	I
-0.8	-0.7	-0.4	1.8	1.9	2.5	3.7	3.8	4.2	4.5

Table 2.2: Hydrophocity scale as given in [KD82] for every of the 20 amino acids, high value corresponding to hydrophobic amino acids.

is the highest score achieved by a segment finishing at position n. The supremum of this process defines the local score $\boldsymbol{H}$. The local score at position n, H_n, is the highest score reached by a segment ending before position n.

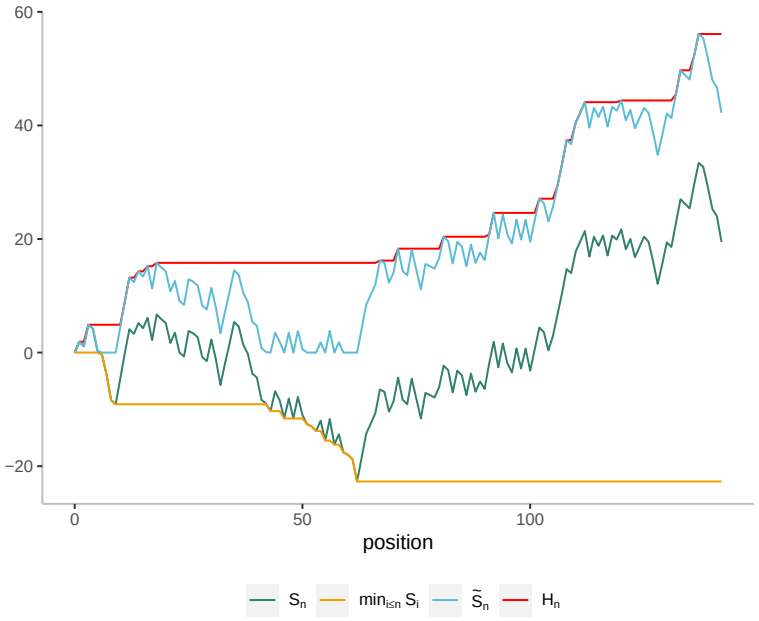


Figure 2.1: Hydrophobicity measure of the Hemoglobin subunit zeta. The partial sums process (S_n) is in green and the corresponding local score process in red. The processes in yellow represents the running minimum of (S_n) while the Lindley process $\tilde{\boldsymbol{S}}$, corresponding to the highest score of a segment finishing at position n is in blue. The segment which achieves the highest score starts at position 61 up to the end of the sequence.

Does this score reveals some organisation in the protein structure? Proposing a statistical test to answering this question implies to specify the distribution of the local score $\boldsymbol{H} = (H_n, n \geq 0)$ under some null hypothesis.

2.1.2 Existing results

When $\boldsymbol{X}$ is a sequence of independent and identically distributed (i.i.d) integer valued random variables, denoting by F_n the cumulative distribution function, Daudin and Mercier

[DM99] proved that

$$F_n(a-1) = P_0^\intercal \Pi^n P_a, \quad \forall n \geq 1, \forall a \in \mathbb{N}^\star, \tag{Eq 2.1.4}$$

where F_n stands for the cumulative distribution function, Π^n is a transition matrix of size $a+1$ elevated to the n^{th} power $P_0 = (1,\, 0,\, \ldots, 0)^\intercal$, and $P_a = (0,\, 0,\, \ldots, 1)^\intercal$. In practice, this result is only computationally available if n and a are not too large.

When $\boldsymbol{X}$ is a sequence of i.i.d. rv's with $\mathbf{E}\{X_i\} < 0$, Karlin, Dembo, and Kawabata [KDK90] have investigated the asymptotic behavior of $\boldsymbol{H}$ and proved that

$$F_n\left(\frac{\log n}{\lambda} + x\right) \underset{n\to\infty}{\approx} \exp\left(-K^\star e^{-\lambda x}\right), \tag{Eq 2.1.5}$$

where $K^\star$ and λ depend only on the probability distribution of X_i. Karlin and Dembo [KD92] proved that Equation Eq 2.1.5 still holds if $\boldsymbol{X}$ is an irreducible aperiodic Markov chain.

In the case $\mathbf{E}\{X_i\} > 0$, the behavior of $\boldsymbol{H}$ is drastically different and

$$H_n \underset{n\to\infty}{\approx} \mathbf{E}\{X_i\}\, n.$$

The study of the phase transition around $\mathbf{E}\{X_i\} = 0$ was the beginning of my research journey.

2.1.3 Study of the case $\mathbf{E}\{X_i\} = 0$

The convergence of the process of partial sums $\boldsymbol{S} = (S_n, n \geq 0)$ has been largely studied under different assumptions regarding $\boldsymbol{X}$. One central result for the development presented in this document is the Donsker Theorem [Bil13]. This theorem proves the convergence in distribution of the linear process $\boldsymbol{S^{(n)}}$ defined by $S^{(n)}(k/n) = (S_k - n\mathbf{E}\{X_1\})/\sqrt{n}$ to a standard Brownian motion, but few results were known regarding the convergence of the local score.

If $\mathbf{E}\{X_i\} = 0$, following the Donsker Theorem, the process of partial sums correctly renormalized converges in distribution to a standard Brownian motion $\boldsymbol{W}$. As a consequence and as suggested in Equation Eq 2.1.3, the corresponding normalized Lindley process tends to $(W_s - \min_{1\leq u\leq s} W_u)$ and thanks to the Paul Levy theorem [RY13, chap II, thm 2.3] which states that $(W_s - \min_{1\leq u\leq s} W_s, s \geq 0) = (|W_s|, s \geq 0)$, we proved in [**DEV03**], that

$$\frac{H_n}{\sqrt{n}} \underset{n\to\infty}{\xrightarrow{(d)}} \sigma W_1^\star, \tag{Eq 2.1.6}$$

where $W_1^\star := \max_{0\leq u\leq 1} |W_u|$.

The cumulative distribution function (cdf) of $\frac{H_n}{\sqrt{n}}$ may be approximated by the cdf of $W_1^\star$, which is given as a series expansion

$$\mathbf{P}\{W_1^\star > x\} = \frac{2}{\pi} \sum_{k\in\mathbb{Z}} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{(2k+1)^2\pi^2}{8x^2}\right\}, \quad x \geq 0. \tag{Eq 2.1.7}$$

As these series converge quickly, from a practical perspective it is acceptable to truncate the series and therefore this result provides an explicit solution to assess the significance of the local score in the context of centered independent random variables.

In [**EV04**], we established the rate of convergence of the local score to its limit and prove that

$$\left|\mathbf{P}\left\{\frac{H_n}{\sigma\sqrt{n}} \geq x\right\} - \mathbf{P}\left\{W_1^\star \geq x\right\}\right| \leq C\sqrt{\frac{\log n}{n}}. \tag{Eq 2.1.8}$$

2.1.4 Study of $\mathbf{E}\{X_i\}$ close to 0.

In order to investigate the behavior around $\mathbf{E}\{X_i\} = 0$, we considered a family $\left\{(X_k^{(N)})_{k\geq 1}, n \geq 1\right\}$ and assume that

$$\lim_{N\to\infty} \sqrt{N}\mathbf{E}\left\{X_i^{(N)}\right\} = \delta \in \mathbb{R}, \quad \lim_{N\to\infty} \mathbf{V}ar\left\{X_i^{(N)}\right\} = \sigma^2 > 0.$$

As an example, we can think to $\boldsymbol{X}$ as a sequence of Bernoulli independent variables with parameter $\frac{1}{2}\left(1 - \frac{1}{\sqrt{n}}\right)$.

In this context, we expect the sequence of partial sums $\boldsymbol{S}$ correctly renormalized to converge to a Brownian motion with drift and in [**DEV03**] we proved that asymptotically the local score behaves like its continuous time counterpart, i.e.,

$$\frac{H_n}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \sigma\xi_{\delta/\sigma}, \tag{Eq 2.1.9}$$

where $\xi_{\delta/\sigma} := \max_{0\leq u\leq 1}\{W_u + \gamma u - \min_{0\leq s\leq u}(W_s + \gamma s)\}$.

As the distribution of ξ_γ was not part of the classical Brownian functional distributions, we investigated its properties and were able to exhibit some equivalent for the tail of the distribution:

$$\mathbf{P}\{\xi_\gamma \geq x\} \underset{x\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\frac{1}{x}e^{-(\gamma-x)^2/2} \tag{Eq 2.1.10}$$

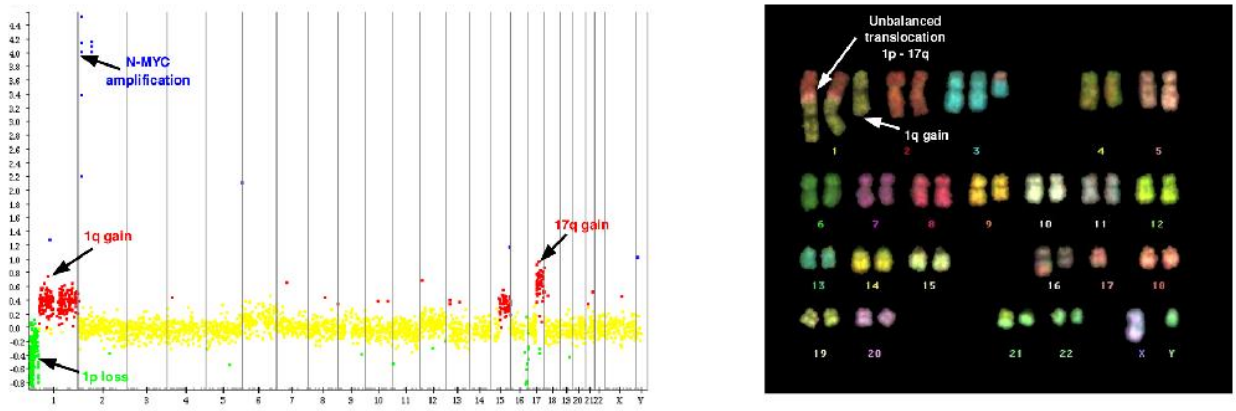
The subject of the local score has been widely studied since those results. We might refer to [Mer18] for an extensive review of known results regarding the properties of the local score and their practical use.

In [**Eti02**], the potential of the local score approach has been illustrated with the detection of segment of proteins with high hydrophobic potential and as mentioned in the introduction of this section, such segment is useful to predict the protein structure (see [Pan+07] for an example).

Recently, the question of the detection of atypical segment came back to me in the context of genomic alterations.

2.2 Detecting shared atypical behavior among individuals

Genomics is a domain where technology is of huge importance and new technologies produce new types of data which allow to explore new problems. The Human Genome Project which was declared complete in 2003 generated a library of cloned DNA fragments and popularize the use of Array-based Comparative Genomic Hybridization (aCGH), a technique that aims to detect chromosomal aberrations on a genomic scale in a single experiment. aCGH microarrays have been developed as genome-wide assays for DNA copy number alterations using the property that fluorescence intensity is related to DNA copy number [Pin+98]. A typical representation of a aCGH profile is depicted in figure 2.2.



(a) IMR32 aCGH profile (b) IMR32 karyotype

Figure 2.2: IMR32 neuroblastoma cell line. On figure a) is the aCGH profile and the corresponding karyotype on figure b).The imbalanced translocation (exchange between chromosomes) between chromosome 1 and chromosome 17 is detected by the aCHG profile which highlights a loss of genetic material at the beginning of chromosome 1 in green, followed by an excess at the end of the chromosome in red. A normal amount of genetic material is materialized by the color yellow. The figure is extracted from [Hup08].

In tumors for example, tumor suppressor genes might be inactivated by deletion while oncogenes are activated by duplication. The study of such changes in comparison with the underlying "normal" genomic state, known as copy number variations (CNVs) provide information related to genome portion affected by a disease. Those studies are important for many types of cancers [Hup08] and have been widely developed in the last decades. Many mathematical models and many efficient algorithms have been and are still being proposed

to infer the copy number alterations from the aCGh profile [Pic+07; TW07; Hoc+13; FR18]. In this thesis, I assume that this treatment of aCGH profiles has been done and I consider a simplified version of such genomic profiles where the sequence $\boldsymbol{L}$ is considered as a sequence of rv's with values in a two letters alphabet $\mathscr{A} = \{alt,\ norm\}$ and L_i represents the character altered or normal at screening position i. The question is now to detect portions of the genome which are more altered than expected.

2.2.1 The local score approach

A first acceptance of the terminology more altered than expected might refer to a portion of genome where the proportion of altered positions is higher than expected in a normal cell. This question might be addressed using a scoring scheme s such that $s(alt) = 1$ and $s(norm) = -1$ and use the local score to identify atypical segment. Consequently, $\mathbf{E}\{X_i\} < 0$ would be a reasonable assumption as the copy numbers are expected to be mostly normal. If we assume that $\boldsymbol{X}$ is either a sequence of iid rv's or a Markov chain under the H_0 hypothesis, the results of the previous section regarding the local score provide the exact distribution (Eq 2.1.4) or the asymptotic distribution (Eq 2.1.5). As a consequence, for every patient, we can identify a portion of genome where the proportion of altered copy number is significantly higher than expected. As each patient is considered independently, the atypical regions may highly differ for each patient.

A second acceptance of the terminology more altered than expected refers to a portion of genome altered in a large proportion of patients as illustrated in Figure 2.2.1 around positions 100. Such alterations are named recurrent alterations. A deviation from normal behavior might occur because a short segment is altered in a large number of patients or because a small proportion of the cohort exhibit the same long alteration. The study of such recurrent alteration is the one of interest in the context of cancer association studies.

Denoting by $\boldsymbol{X}^j$ the scored profile of patient j, the significance of the event "a portion of length ℓ is altered in a patients" might be explored through the study of the cumulative profile $\boldsymbol{Y}^{(N)} = \sum_{j=1}^{N} \boldsymbol{X}^{(j)}$. Let's denote by A_i the number of altered profiles at position i, $N - A_i$ being the number of normal profiles. The sequence $\boldsymbol{Y}^{(N)}$ counts, at each position i the exceed of patients with altered loci compared to the number of patients with normal genome $Y_i = A_i - (N - A_i) = 2A_i - N$. The value at position k of the corresponding cumulative sum $S_k = \sum_{i=1}^{k} Y_i$, with $S_0 = 0$, as defined in equation Eq 2.1.1 will count the exceed of altered positions among the N patients up to position k: the event $\{S_k > 0\}$ corresponds to a situation where the number of altered positions, when considering all loci up to position k for the N patients, exceeds the number of normal loci. In this context the local score will correspond to the longest portion of the genome where the number of altered loci exceeds the number of normal loci.

If the individual sequence $\boldsymbol{X}^j$ are i.i.d rv's sequences, so is the cumulative profile $\boldsymbol{Y}^{(N)}$. If the individual sequence $\boldsymbol{X}^i$ are Markov chain it is easy to prove that so is the cumulative profile $\boldsymbol{Y}^{(N)}$.

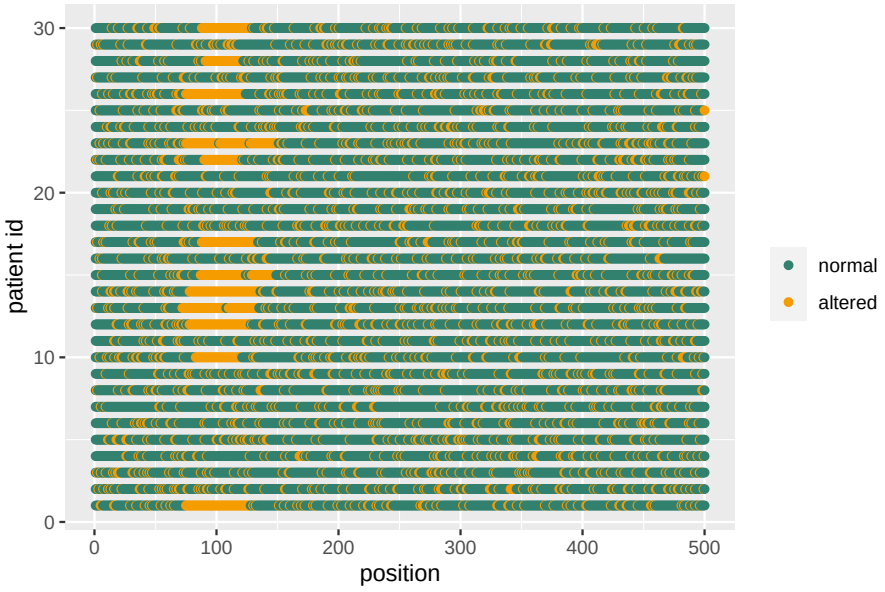


Figure 2.3: Illustration of aCGH profiles for a cohort of patients. Altered loci are represented in yellow and normal loci are represented in green blue. The genome portion around position 100 appears to be more frequently altered in a large number of patients than the entire genome.

Lemma 2.1. *Let $(\boldsymbol{X}^i)_{1\leq i\leq N}$ a set of independent homogeneous Markov chains with values in $\{-1,1\}$ with the same transition matrix*

$$\Pi = \begin{pmatrix} 1-\Pi_{-1,1} & \Pi_{-1,1} \\ \Pi_{1,-1} & 1-\Pi_{1,-1} \end{pmatrix}$$

and the same initial distribution $\nu_1 = \mathbf{P}\{X_1^i = 1\}$. Let $\boldsymbol{Y}^{(N)} := \sum_{j=1}^N \boldsymbol{X}^j$, then $\boldsymbol{Y}^{(N)}$ is a homogeneous Markov Chain with values in $\mathscr{E} = \{-N+2k;\ k=0,\dots,N\}$ and a transition matrix Π^+ such that, for any $(l,m)\in\mathscr{E}^2$,

$$\Pi^+_{ml} = \mathbf{P}\{Y_i = l | Y_{i-1} = m\} = \mathbf{P}\{Z_{1m} + Z_{2m} = (N+l)/2\},\ \textit{with } Z_{1m} \perp\!\!\!\perp Z_{2m},$$
$$Z_{1m} \sim \mathscr{B}\left(\frac{N+m}{2}, \Pi_{-1,1}\right) \textit{ and } Z_{2m} \sim \mathscr{B}\left(\frac{N-m}{2}, \Pi_{1,-1}\right) \tag{Eq 2.2.11}$$

and initial distribution ν^+, such that

$$\nu^+_l = \mathbf{P}\{Y_1 = l\} = \mathbf{P}\{Z_1 = (l+N)/2\}, \qquad \textit{with } Z_1 \sim \mathscr{B}(N, \nu_1).$$

Proof. First notice that for any $(m,l)\in\mathscr{E}^2$, $N-l$, $N+l$, $m-l$ and $m+l$ are always even and that the event $\{Y_i = l\}$ equals to the event $\{A_i = (l+N)/2\}$, i.e. the event of having $(l+N)/2$ chains among the N in state 1. Therefore

$$\nu^+_l = \binom{N}{(m+N)/2} \nu_1^{(m+N)/2} (1-\nu_1)^{(N-m)/2}.$$

The term Π^{+}_{ml} of the transition matrix might be reformulated in terms of number of altered profiles :

$$\begin{aligned}\Pi^{+}_{ml} &= \mathbf{P}\left\{Y_{k+1} = l | Y_k = m\right\} \\ &= \mathbf{P}\left\{A_{k+1} = (l+N)/2 | A_k = (m+N)/2\right\} \\ &= \sum_{s=0\vee(m-l)/2}^{(N-l)/2\wedge(N+m)/2} \binom{(N+m)/2}{s} \Pi^{s}_{-1,1} \Pi^{(N+m)/2-s}_{-1,-1} \binom{(N-m)/2}{(l-m)/2+s} \Pi^{(l-m)/2+s}_{1,-1} \Pi^{(N-m)/2+s}_{1,1}\end{aligned}$$

■

The number of altered positions is expected to be smaller than the number of normal positions, therefore $\mathbf{E}\left\{Y_i\right\} < 0$ is a reasonable assumption.

If individual profiles are assumed to be sequence of independant rv's or a Markov chain, the exact distribution of the local score for the cumulative profile is given by the results in [DM99] and an asymptotic approximation is available in [KD92]. Using the local score approach, we are able to identify the longest portion of the sequence where the proportion of altered position exceeds 0.5. This local score approach is highly dependent on the choice of the scoring scheme. According to the proposed scoring scheme defined above, an atypical segment designs a segment where the proportion of altered loci among all patients exceed 0.5. Going for other scoring scheme would lead to different definition of atypicality and there is very few clues for a choice of a relevant scoring scheme.

Robin and Stefanov [RS09] and Robin and Stefanov [RS15] proposed to model the aCGH profile of a single patient by a sequence of random variable with values in $\{0, 1\}$. Although this approach could be thought as just an alternative simple scoring scheme where $s(alt) = 1$ and $s(norm) = 0$, it does not depend on this choice and as so propose an interesting alternative.

2.2.2 Previous works

Small sequences and few patients : Markov Chains approach

In [RS09][2], the authors propose to model an aCHG profile as a two states Markov chain as presented previously except that the two states are denoted by 0 and 1 where 0 refers to normal state whereas 1 stands for abnormal state. They focus in simultaneous occurrences of runs of 1's with length ℓ. Defining an ad-hoc Markov Chain on a larger state space, they are able to bound the probability to observe at least M simultaneous runs of length ℓ. The complexity of this approach depends on the number of positions in the profiles.

[2]In order to unify the presentation, the notations I used in the present document differ from the original paper.

Long sequences and few patients : Continuous time process and renewal theory

With the emergence of high speed sequencing, the length of the profiles tend to approach the million of positions and the evaluation of the previous probability raises serious computational problems.

To circumvent those computational issues, the same authors proposes in [RS15] to use a continuous time process and model a patient profile $\boldsymbol{X}^{(i)} = (X_t^{(i)}, 0 \leq t \leq T)$ as a 2-state continuous time Markov process over the interval $[0, T]$, characterized by its infinitesimal generator

$$Q_X = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

They define a recurrent alteration as a significantly long alteration shared by a significant number of patients. The cumulative process $\boldsymbol{Y}^{(N)}$ is defined by $\boldsymbol{Y}^{(N)} = \sum_{i=1}^{N} \boldsymbol{X}^i$, $\boldsymbol{X}^i$ being independent continuous time Markov processes. As the processes $\boldsymbol{X}^i$ are independent, the process $\boldsymbol{Y}^{(N)}$ is a birth and death process with state space $\mathscr{S}^N := \{0, \cdots, N\}$ and transition intensities $\lambda_i = (N - i)\lambda$ (from i to $i + 1$) and intensities $\mu_i = i\mu$ (from i to $i - 1$). Figure 2.4 depicts an example of N profiles with the corresponding cumulative profile.

The significance level of a recurrent alteration of length ℓ shared by at least m patients is bounded by $\mathbf{P}\left\{A_\ell^{(N)}\right\}$ with

$$A_\ell^{(N)} = \left\{\exists \tau \in [0, T - \ell] : \forall t \in [\tau, \tau + \ell], Y_t^{(N)} \geq m\right\}, \qquad \text{(Eq 2.2.12)}$$

and the authors propose a solution to derive this probability, which relies on the inversion of a Laplace transform and is intractable for large N.

2.2.3 Long sequences and large cohort

This section presents an on going collaboration with Laurent Decreusefond, Gabriel Lang, Stéphane Robin and Pierre Vallois as an attempt to address the question for large cohort. The first motivation for this works has been presented above. Our second motivation comes from telecommunications and has been brought by Laurent Decreusefond. It concerns the functioning of operated systems like GSM or 4G which strongly depends on the signal to interference ratio (SIR).

As proposed in [RS15], the process $\boldsymbol{Y}^{(N)}$ of interest is a continuous-time Markov process with state space $\mathscr{S}^N$ and we are interested in characterizing the distribution of its longest time spent above a given threshold m within a time period ℓ for large N. Again, the strategy to avoid overly burdensome calculations will be to consider the continuous time, continuous space limit process.

In this document, I will omit the very technical aspects of the study (those aspects are detailed in a publication in preparation Decreusefond et al. [**Dec+212**]) and present only the key steps of the proposed approach. We restrict ourselves to the case $\mu = \lambda$. In [RS15], the process $\boldsymbol{Y}^{(N)} = (Y^{(N)}(t), 0 \leq t \leq 1)$ is proved to be a continuous time Markov Chain, and because of the assumption $\mu = \lambda$, the jump process is a Poisson process with intensity

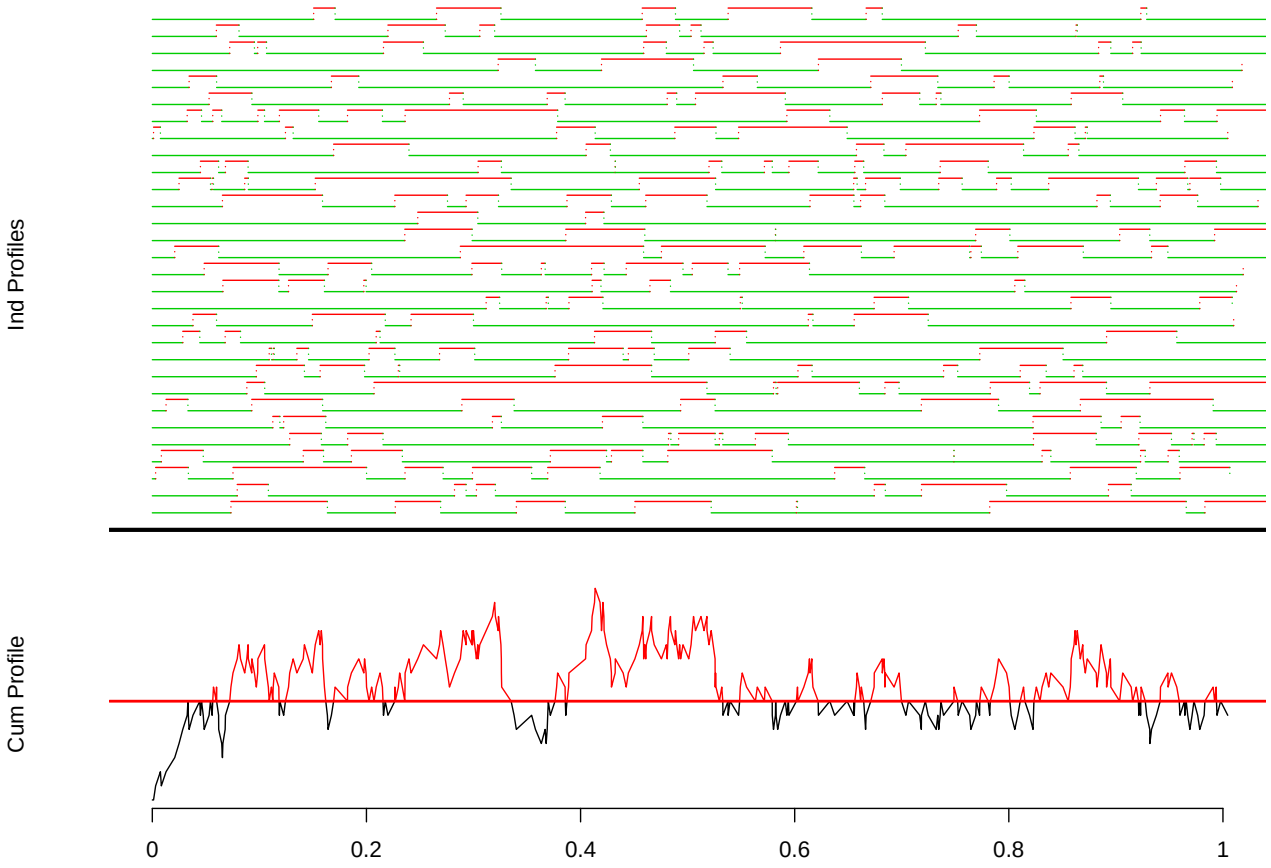


Figure 2.4: Illustration of N realizations of 2 states Markov process (state 0 in green and state 1 in red) and the corresponding cumulative profile. The excursion above a threshold M are colored in red.

λN, i.e. the sojourn times of $\boldsymbol{Y}^{(N)}$ are independent and follow an exponential distribution with expectation equals $(\lambda N)^{-1}$.

The quantity of interest given by Equation Eq 2.2.12 can be reformulated in terms of excursions above a given threshold m. For any $t \geq T$, let's define $l(\boldsymbol{Y}^{(N)}, m, t)$ the last instant s before t such that $Y_0^{(N)}(s) = m$. Symmetrically let's define $r(\boldsymbol{Y}^{(N)}, m, t)$, as the first instant s after t where $Y_0^{(N)}(s) = m$. More formally

$$l(\boldsymbol{Y}^{(N)}, m, t) = \begin{cases} 0 \text{ if } \left\{s : s < t, Y_0^{(N)}(s) = m\right\} = \emptyset, \\ \sup\left\{s < t : Y_0^{(N)}(s) = m\right\}, \text{otherwise,} \end{cases}$$

$$r(\boldsymbol{Y}^{(N)}, m, t) = \begin{cases} 0 \text{ if } \left\{s : t < s < T, Y_0^{(N)}(s) = m\right\} = \emptyset, \\ \inf\left\{s : t < s < T, Y_0^{(N)}(s) = m\right\} \end{cases}$$

For a given t, the path from $l(\boldsymbol{Y}^{(N)}, m, t)$ to $r(\boldsymbol{Y}^{(N)}, m, t)$ is called an excursion away from m. These definitions are illustrated in Figure 2.5. The length of the excursion around t away from m is given by $r(\boldsymbol{Y}^{(N)}, m, t) - l(\boldsymbol{Y}^{(N)}, m, t)$. It is clear that the event of interest $A_\ell^{(N)}$, defined in Equation Eq 2.2.12, verifies:

$$A_\ell^{(N)} = \left\{ \sup_{t \in [0,T]} \left\{ r(\boldsymbol{Y}^{(N)}, m, t) - l(\boldsymbol{Y}^{(N)}, m, t) : Y_0^{(N)}(t) > m \right\} > \ell \right\}.$$

If the size of the cohort N goes to infinity, we are lead to study the asymptotic distribution of the lengths of excursions above m of the process $\boldsymbol{Y}^{(N)}$ when N goes to infinity.

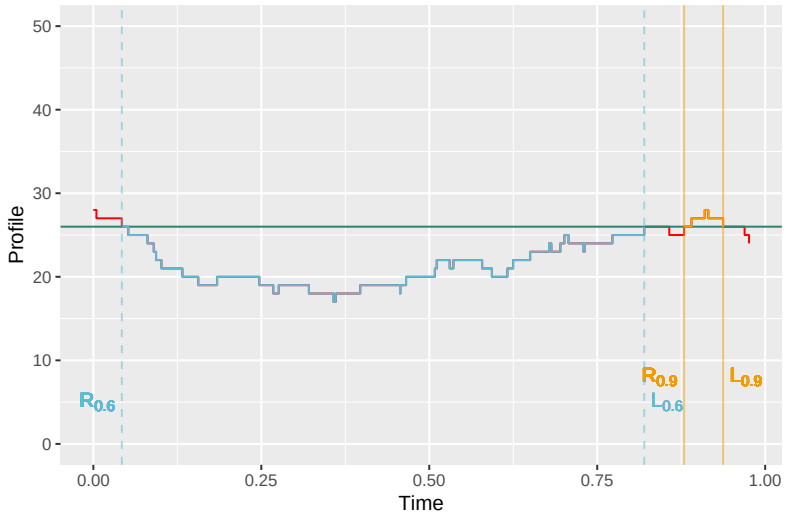


Figure 2.5: Illustration of two excursions from $m = 26$ and the corresponding values of the processes $l(\boldsymbol{Y}^{(N)}, m, t)$ and $r(\boldsymbol{Y}^{(N)}, m, t)$ for $t = 0.6$ and $t = 0.9$.

Let's consider $\boldsymbol{U}^{(N)}$ a centered and scaled version of $\boldsymbol{Y}^{(N)}$:

$$\boldsymbol{U}^{(N)} := \frac{\boldsymbol{Y}^{(N)} - N/2}{\sqrt{N/4}} = \frac{\sum_i (\boldsymbol{X}^i - 1/2)}{\sqrt{N/4}}. \tag{Eq 2.2.13}$$

The event of interest A_ℓ^N defined in Equation Eq 2.2.12 can be reformulated as

$$A_\ell^N = \left\{ \exists \tau \in [0, T-\ell] : \forall t \in [\tau, \tau+\ell], U^{(N)}(t) \geq a \right\}.$$

Proposition 2.1.1. *$\boldsymbol{U}^{(N)}$ converges in distribution to $\boldsymbol{U}_\lambda$, a standard stationary Ornstein Uhlenbeck process defined by $U(0) \sim \mathcal{N}(0,1)$ and the stochastic differential equation:*

$$dU_\lambda(t) = -2\lambda U_\lambda(t)dt + \sqrt{4\lambda} dW(t),$$

where $\boldsymbol{W} = (W(t), t \geq 0)$ stands for the standard Brownian motion.

Proof. Robin and Stefanov [RS15] have proved that $\boldsymbol{Y}^{(N)}$ is a continuous time Markov chain and so is $\boldsymbol{U}^{(N)}$ with space state $\mathscr{S}^N = u \in \left\{ -\sqrt{N} + \frac{2k}{\sqrt{N}},\ k = 0, \dots N \right\}$. Let Q^N be the infinitesimal generator of $\boldsymbol{U}^{(N)}$,

$$Q^N = \begin{pmatrix} -\lambda N & \lambda N & 0 & \dots & \dots & \dots & 0 \\ \lambda & -\lambda N & \lambda(N-1) & \ddots & & & \vdots \\ 0 & 2\lambda & -\lambda N & \lambda(N-2) & \ddots & & \vdots \\ \vdots & \ddots & & & & \ddots & \vdots \\ \vdots & & \ddots & & & & 0 \\ & & & \ddots & \lambda(N-1) & -\lambda N & \lambda \\ 0 & \dots & \dots & \dots & 0 & \lambda N & -\lambda N \end{pmatrix}.$$

The proof is then a direct application of Ethier and Kurtz [EK86, thm 4.1, p.354]. ■

Let us consider A_ℓ, the analog of A_ℓ^N for the limit process $\boldsymbol{U}$. Thanks to the continuity of the process $\boldsymbol{U}$, we are able to prove that

$$A_\ell = \left\{ \sup_{0 \leq s \leq T-l} \inf_{s \leq u \leq s+\ell} U(s) > m \right\}.$$

and since A_ℓ is expressed through continuous functional, we are finally able to prove that

$$\mathbf{P}\left\{ A_\ell^{(N)} \right\} \underset{N \to \infty}{\longrightarrow} \mathbf{P}\left\{ A_\ell \right\},$$

The probability of the event, there exists a segment of length at least ℓ, for which at least M patients exhibit an alteration, can be approximated by the probability that the length of the longest excursion of standard stationary OU process exceeds ℓ.

2.2.4 The distribution of the longest excursion of an Ornstein Uhlenbeck process

Pitman and Yor [PY972] and Pitman and Yor [PY92] studied the laws of the excursions away from 0 for a Brownian motion and a centered Ornstein Uhlenbeck and proposed a very straightforward methods to sample them. However, to the best of our knowledge, nothing is known regarding the excursions of an OU process away from m with $m \neq 0$. We propose to develop an importance sampling algorithm where the proposals are based on the simulation of Brownian motion excursion.

The quantity of interest $\mathbf{P}\{A_\ell\}$ can be expressed as $\mathbf{E}\{F(L(\boldsymbol{U}_\lambda, m, 1))\}$, where $L(\omega, m, t)$ stands for the length of the longest excursion of the process ω above m before time t and F is the suitable indicator function. The proposed algorithm makes use of three key ingredients

- Let σ_m be the first time a stationary OU process $\boldsymbol{U}_\lambda$ hits m, $\sigma_m := \inf_{t\geq 0}\{U_\lambda(t) > m\}$. Conditionally on $\{\sigma_m > 1, U_\lambda(0) < m\}$, $F(L(\boldsymbol{U}_\lambda, m, t)) = 0$, and conditionally on $\{\sigma_m > 1, U_\lambda(0) > m\}$, $F(L(\boldsymbol{U}_\lambda, m, t)) = t$.
  On $\{\sigma_m \leq 1\}$, thanks to the strong Markov property $\widetilde{\boldsymbol{U}}_\lambda^{-m,0} = (U(t+\sigma_m) - m, 0 \leq t \leq 1 - \sigma_m)$ is an OU process starting from 0 centered on $-m$ whose excursions away from 0 we want to study. Figure 2.6 illustrates this change in time and space.

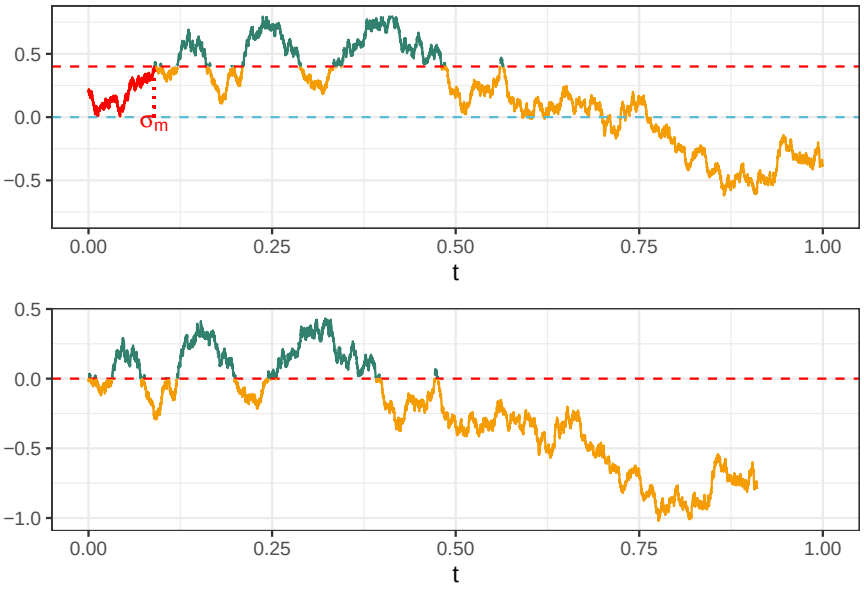


Figure 2.6: The top panel is a realization of $\boldsymbol{U}_\lambda$. The portion in red corresponds to the process before σ_m, while above (reps. below) m excursions are in green (resp. in yellow). The final portion of the process, does not reach m and is called the meander. The bottom panel is corresponding realization of $\widetilde{\boldsymbol{U}}_\lambda^{-m,0}$.

- Conditionally on $\sigma_m < 1$,

$$\begin{aligned}\mathbf{E}\left\{F\left(L(\boldsymbol{U}_\lambda, m, 1)\right)|\sigma_m\right\} &= \mathbf{E}\left\{F\left(L(\widetilde{\boldsymbol{U}}_\lambda^{-m,0}, 0, 1-\sigma_m)\right)\right\} \\ &= \mathbf{E}_Q\left\{F\left(L(\boldsymbol{W}, 0, 1-\sigma_m)\right)\Lambda(1-\sigma_a)|\sigma_m\right\},\end{aligned}$$

  where Λ is the Radon Nikodym derivative of the initial measure with respect to the Wiener measure, $\Lambda(t) = \exp\left(m\tau W(t) + \frac{\tau}{2}W(t)^2 - \frac{\tau t}{2} - \frac{\tau^2}{2}\int_0^t (m + W(s))^2\,\mathrm{d}s\right)$.

- Thanks to the representation of a Brownian motion in terms of its excursions and the results given by Pitman and Yor [PY82] the quantity Λ can be expressed in terms of the length and the sign of the excursions of a Brownian motion and an additional term which involves a numerical integration of the Brownian meander.

- Devroye [Dev10] provides efficient simulation methods for the Brownian meander simulation.

The main idea of the algorithm, would be to sample σ_m, $U_\lambda(0)$, the K longest excursions of a Brownian motion and their signs thanks to the method proposed by Pitman and Yor [PY972]. The result of the algorithm would be the longest excursion associated with a positive sign and its weight. As mentioned before, this work is still in progress and the practical implementation of the algorithm has now to be explored. This point will be discussed in the final conclusion of this document.

2.3 Conclusions

The results obtained or the directions to be explored presented in this chapter are good examples of applied questions that have raised interesting theoretical questions.

As I mentioned in the introduction, I am committed to producing results that are of practical interest and a limit approximation is of poor interest without the corresponding rate of convergence. This aspect has been explored for the local score of the first section and is a question I would like to explore before publishing the results described in section 2.2.3. Thanks to an integral representation of the Ornstein Uhlenbeck process in terms of of martingale difference arrays and the results obtained by Kubilius [Kub94], we are now working to prove that the rate of convergence equals $N^{-1/4-\varepsilon}\log(N)$.

There is still work to go to propose an efficient algorithm to compute the significance of a recurrent alteration for large cohort and the different possible directions are presented in the perspectives section of the conclusion of this document. It is worth noting that Robin and Stefanov [RS15] proposed a continuous-time approach when their initial method [RS09] showed its combinatorial limitation. Again, our solution has been to move to continuous processes to tackle the issue for large cohorts. This approach opens the door to an algorithm whose complexity no longer depends on the size of the cohort.

In the next chapter I also present discrete and continuous approaches in a completely different domain and again I illustrate that, while the probabilistic objects involved in continuous approaches may appear more difficult at first glance, they often lead to simplifications from a computational point of view.

Chapter bibliography

[Alt+90] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[Bil13] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[**Dec+212**] Laurent Decreusefond, Marie-Pierre Etienne, Gabriel Lang, Stéphane Robin, and Pierre Vallois. "Convergence of the sum of pure jump Markov unitary processes to an Ornstein Uhlenbeck process". 2021.

[**DEV03**] Jean-Jacques Daudin, Marie-Pierre Etienne, and Pierre Vallois. "Asymptotic behavior of the local score of independent and identically distributed random sequences". In: *Stochastic Processes and their Applications* 107.1 (Sept. 2003).

[Dev10] Luc Devroye. "On exact simulation algorithms for some distributions related to Brownian motion and Brownian meanders". In: *Recent Developments in Applied Probability and Statistics*. Springer, 2010, pp. 1–35.

[DM99] Jean-Jacques Daudin and Sabine Mercier. "Distribution exacte du score local d'une suite de variables indépendentes et identiquement distribuées". In: *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* 329.9 (1999), pp. 815–820.

[EK86] Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

[**Eti02**] Marie-Pierre Etienne. "Le score local: un outil pour l'analyse de séquences biologiques". PhD thesis. Université Henri Poincaré-Nancy 1, 2002.

[**EV04**] Marie-Pierre Etienne and Pierre Vallois. "Approximation of the distribution of the supremum of a centred random walk. Application to the local score". In: *Methodology and Computing in Applied Probability* 6.3 (2004), pp. 255–275.

[FM06] Martin J Fischer and Denise M Bevilacqua Masi. "Analyzing internet packet traces using Lindley's Recursion". In: *Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference. 2006, pp. 2195–2201.

[FR18] Paul Fearnhead and Guillem Rigaill. "Changepoint detection in the presence of outliers". In: *Journal of the American Statistical Association* (2018), pp. 1–15.

[Hoc+13] Toby Dylan Hocking, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Valentina Boeva, Julie Cappo, Olivier Delattre, Francis Bach, and Jean-Philippe Vert. "Learning smoothing models of copy number profiles using breakpoint annotations". In: *BMC bioinformatics* 14.1 (2013), p. 164.

[Hup08] Philippe Hupé. "Biostatistical algorithms for omics data in oncology-Application to DNA copy number microarray experiments". PhD thesis. AgroParisTech, 2008.

[Igl+72] Donald L Iglehart et al. "Extreme values in the GI/G/1 queue". In: *The Annals of Mathematical Statistics* 43.2 (1972), pp. 627–635.

[KD82] Jack Kyte and Russell F Doolittle. "A simple method for displaying the hydropathic character of a protein". In: *Journal of Molecular Biology* 157.1 (1982), pp. 105–132.

[KD92] Samuel Karlin and Amir Dembo. "Limit distributions of maximal segmental score among Markov-dependent partial sums". In: *Advances in Applied Probability* 24.1 (1992), pp. 113–140.

[KDK90] Samuel Karlin, Amir Dembo, and Tsutomu Kawabata. "Statistical composition of high-scoring segments from molecular sequences". In: *The Annals of Statistics* 18.2 (1990), pp. 571–581.

[Kub94] K Kubilius. "Rate of convergence in the invariance principle for martingale difference arrays". In: *Lithuanian Mathematical Journal* 34.4 (1994), pp. 383–392.

[MD01] Sabine Mercier and Jean-Jacques Daudin. "Exact distribution for the local score of one iid random sequence". In: *Journal of Computational Biology* 8.4 (2001), pp. 373–380.

[Mer18] Sabine Mercier. "Distribution du score local pour la détection de régions atypiques au sein de séquences". PhD thesis. Université Toulouse III Paul Sabatier (UT3 Paul Sabatier), 2018.

[Pan+07] Chi NI Pang, Kuang Lin, Merridee A Wouters, Jaap Heringa, and Richard A George. "Identifying foldable regions in protein sequence from the hydrophobic signal". In: *Nucleic acids research* 36.2 (2007), pp. 578–588.

[Pic+07] Franck Picard, Stéphane Robin, E Lebarbier, and J-J Daudin. "A segmentation/clustering model for the analysis of array CGH data". In: *Biometrics* 63.3 (2007), pp. 758–766.

[Pin+98] Daniel Pinkel, Richard Segraves, Damir Sudar, Steven Clark, Ian Poole, David Kowbel, Colin Collins, Wen-Lin Kuo, Chira Chen, Ye Zhai, et al. "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". In: *Nature Genetics* 20.2 (1998), p. 207.

[PY82] Jim Pitman and Marc Yor. "A decomposition of Bessel bridges". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 59.4 (1982), pp. 425–457.

[PY92] Jim Pitman and Marc Yor. "Arcsine laws and interval partitions derived from a stable subordinator". In: *Proceedings of the London Mathematical Society* 3.2 (1992), pp. 326–356.

[PY972] James W Pitman and Marc Yor. "On the lengths of excursions of some Markov processes". In: *Séminaire de probabilités de Strasbourg* 31 (1997), pp. 272–286.

[RS09] Stephane Robin and VT Stefanov. "Simultaneous occurrences of runs in independent Markov chains". In: *Methodology and Computing in Applied Probability* 11.2 (2009), pp. 267–275.

[RS15] Stéphane Robin and Valeri T Stefanov. "Detection of significant genomic alterations via simultaneous minimal sojourns at a state by independent continuous-time markov chains". In: *Methodology and Computing in Applied Probability* 17.2 (2015), pp. 479–487.

[RY13] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Vol. 293. Springer Science & Business Media, 2013.

[TW07] Robert Tibshirani and Pei Wang. "Spatial smoothing and hot spot detection for CGH data using the fused lasso". In: *Biostatistics* 9.1 (2007), pp. 18–29.

[Wat95] Michael S Waterman. *Introduction to Computational Biology: Maps, Sequences and Genomes*. CRC Press, 1995.

Chapter 3

Stochastic processes and movement modelling

Contents

As mentioned by Lucas Börger in an introduction of Journal of Animal Ecology, Charles Sutherland Elton, who developed the field of Animal Ecology " *stressed the necessity for the field to distinguish itself from the approaches used by plant ecologists, due to the different principles governing animal systems, most notably as 'animals move about' (Elton 1933). Since then the study of the causes and consequences of movement of organisms has become a central question in ecology, providing a link between individual behavior and spatial processes, from population to community ecology and beyond".*

Those research interests initiate the field of movement ecology. More specifically, the corner stone of movement ecology is the assumption that internal states/behavior of an individual, environmental covariates, and interaction between individuals affect their movement and therefore the study of this movement should provide insights on all those aspects and allow to address some interesting ecological questions.

Nathan et al. [Nat+08] sum up this idea in a conceptual chart of movement ecology which is depicted in figure 3.1. This figure suggests that for each species, the two main drivers of the movement are i) environment and ii)internal states, but little is known on how those drivers modify the movement.

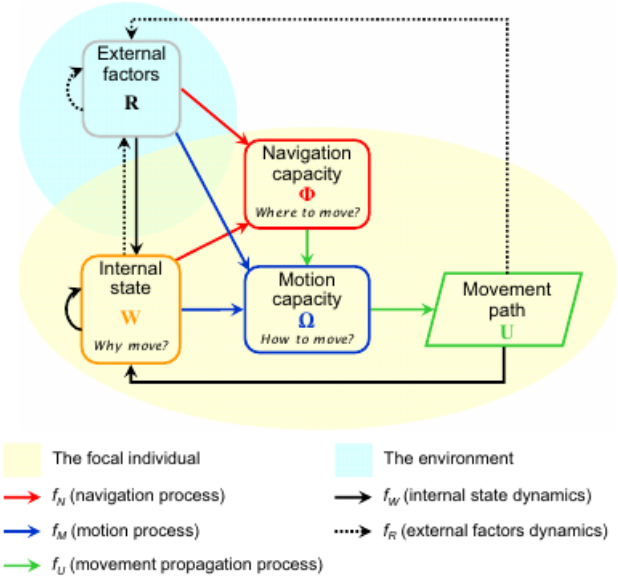


Figure 3.1: Figure from Nathan et al. [Nat+08] to illustrate the underlying processes involved in movement.

This conceptual chart is adequately formalized using a graphical model. A graphical model is a probabilistic model whose conditional (in)dependence structure between random variables is given by a graph, the Directed Acyclic Graph (DAG). Embracing the ideas of hierarchical modeling, Figure 3.2 uses a DAG to provideone possible overview of every component which could or should be included in an integrated movement model.

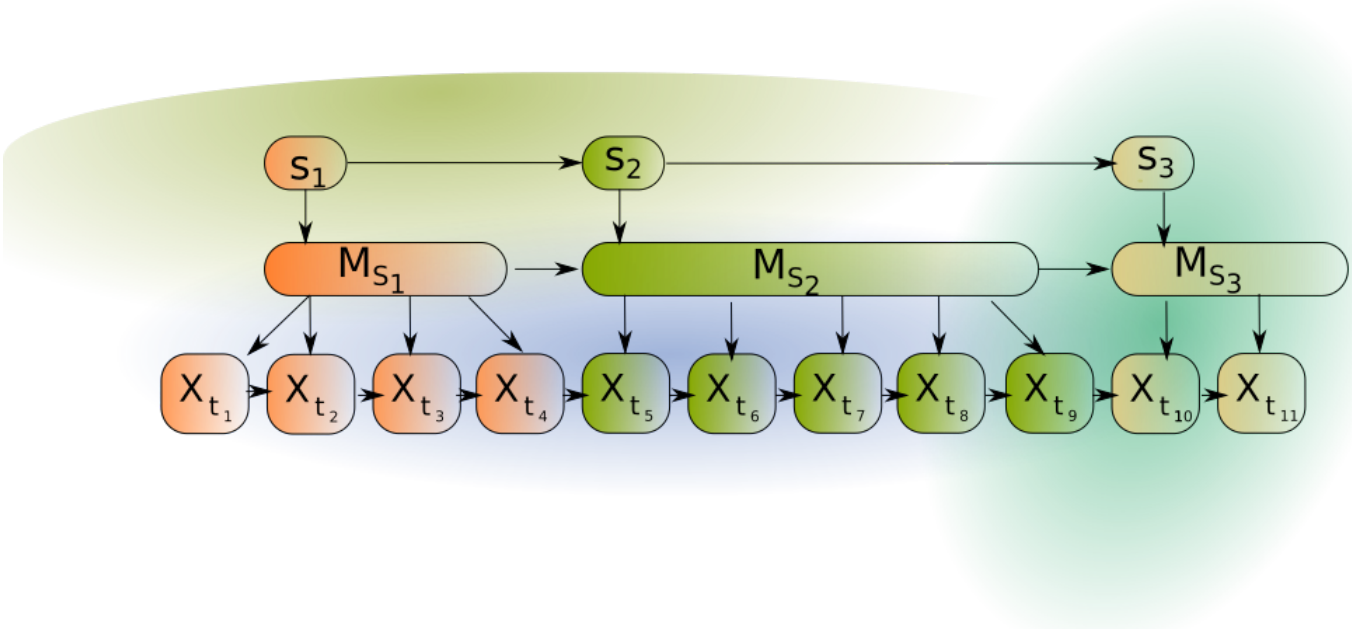


Figure 3.2: Illustration of the different aspects to be accounted for in a modeling approach. $\boldsymbol{S}$ denotes the internal states, M_S stands for the movement model given state S, while X_{t_i} is the observed position at time t_i. The background is the changing environment which might affect every aspects (internal states like opportunistic foraging behavior, transition from one internal state to another, movement itself depending on the attractiveness of the environment, ...).

- At the top of the hierarchy stands the process describing the internal states of the individual $\boldsymbol{S}$[1]. This process is changing over time, with some potential dependence to the past and to the environment.
- The individual movement process is assumed to be driven by this internal state corresponding to a movement model M_S.
- Finally, integrating the sampling process, and conditionally on the movement model, we can define the sequence of observed positions $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$.
- Environmental drivers can affect any of the previous components. The colored changing background in figure 3.2 highlights the idea that the environment might affect every components.

At this point, it should be noted that the sequence $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and the potential covariates are the only available information. Based on these observations, the two main questions of movement ecology I have contributed to are:

[1]This corresponds to the internal state denoted W in yellow in the representation proposed by Nathan et al. [Nat+08].

- Q1 : How individuals use space?
- Q2 : How movement informs us about internal states?

This chapter starts with a brief overview of movement data specifities 3.1.1 and then develops the ecological concepts useful to address Q1 and Q2. Section 3.4 focuses on the recent advances in behavioral reconstruction from path analysis while section 3.3.3 presents how Stochastic Differential Equations provide promising and flexible framework to propose more realistic movement models, potentially including covariates. The last section proposes to combine these two approaches to propose reasonably realistic movement models with switching internal behavior.

3.1 Movement data

Monitoring movement is one specific field of the bio-logging science define as *"the use of miniaturized animal-attached tags for logging and/or relaying of data about an animal's movements, behaviour, physiology and/or environment"* in [RH09]. Bio-logging science has received more and more attention in the recent years[2] and the tags attached to individuals become more and more sophisticated (pressure sensor, accelerometer, physiological sensor, environmental monitoring, on board camera, etc) which raise a number of questions in terms of data storage, data management and data analysis.

My research focuses exclusively on telemetry data. Cagnacci et al. [Cag+10] claim that *"In the history of science, rapid conceptual advances have often been stimulated by technological innovations. Such technological milestones included Galileo and Kepler designing and looking into telescopes, finding evidence for Copernicanism (and falsifying the Tolemaic system); Hooke and van Leeuwenhoek peering into microscopes and describing cells, thus laying the basis for modern microbiology; Sanger and Maxam and Gilbert developing DNA sequencing, which spawned molecular biology; and high-speed computers fostering the emergence of nonlinear dynamics [...] New telemetry technology allows us to monitor and to map the details of animal movement, securing vast quantities of such data even for highly cryptic organisms."* New statistical methods are also required to analyze those data and extract relevant ecological information.

3.1.1 Some technological aspects

From the first studies in the 1960s using VHF transmitter [CL63], telemetry device recover a large variety of technologies and can be classified in archival tags (data are stored on the device which has to be recovered and unloaded) versus transmitting tag (the data are sent to communications satellites). Each technology has to cope with a balance between three different aspects:

[2] A first international Symposium hold in 2003, http://polaris.nipr.ac.jp/~penguin/oogataHP/IndexC.html. The International Bio-Logging Society has been created in 2016, https://www.bio-logging.net/

	event-id	timestamp	location-long	location-lat
1	677436629	2011-06-15 17:35:18	-59.97949	43.92495
2	677436630	2011-06-15 17:50:19	-59.98273	43.92548
3	677436631	2011-06-15 18:05:32	-59.98968	43.92582
4	677436632	2011-06-15 18:21:27	-59.99033	43.92613
5	677436633	2011-06-15 18:36:31	-59.98896	43.92525
6	677436634	2011-06-15 18:51:23	-59.98394	43.92564
7	677436635	2011-06-15 19:06:20	-59.98566	43.92499
8	677436636	2011-06-15 19:22:18	-59.98785	43.92406
9	677436637	2011-06-15 19:37:18	-59.98073	43.92603
10	677436638	2011-06-15 22:27:41	-59.98105	43.92588
11	677436639	2011-06-15 22:48:23	-59.96713	43.93259
12	677436640	2011-06-15 23:03:31	-59.98013	43.92653

Table 3.1: Example of grey seals (*Halichoerus grypus*) equipped with GPS tags on the Scotian Shelf (Atlantic Canada), [LBI15]. The full dataset is available on https://www.datarepository.movebank.orghttps://www.datarepository.movebank.org/handle/10255/move.451.

- Energy consumption,
- Precision,
- Data transmission.

The size of the battery is directly related to the amount of available energy. Therefore small animals can be equipped only with small battery and have limited on-board energy. This energetic constraint limits the number of recorded relocations. Since data transmission is an energy-consuming operation, transmitting tag requires larger battery. Environmental conditions also puts constraints on the device type as transmission to satellite is not possible in an aquatic environment or in in a dense canopy forestrial environment.

In chapter 1 Hooten et al. [Hoo+17] gives a nice overview of different tags and Kays et al. [Kay+15] propose a more detailed review of tags technology in terrestrial environment. Whatever the choice of the technology, the movement data include at least time, latitude, longitude as illustrated in the example of grey seals borrowed from [Bak+15] in table 3.1.

3.1.2 Regularity of the sampling process

In the grey seals example, data appear at first glance to be regularly spaced in time (every 15mn) but there was almost 3 elapsed hours between observations 9 and 10. This lack of regularity is classical especially with marine mammals, since the relocations can only be captured when the animal is out of the water (breathing or resting). In the presented dataset, the median time difference between successive observations vary from 18 seconds

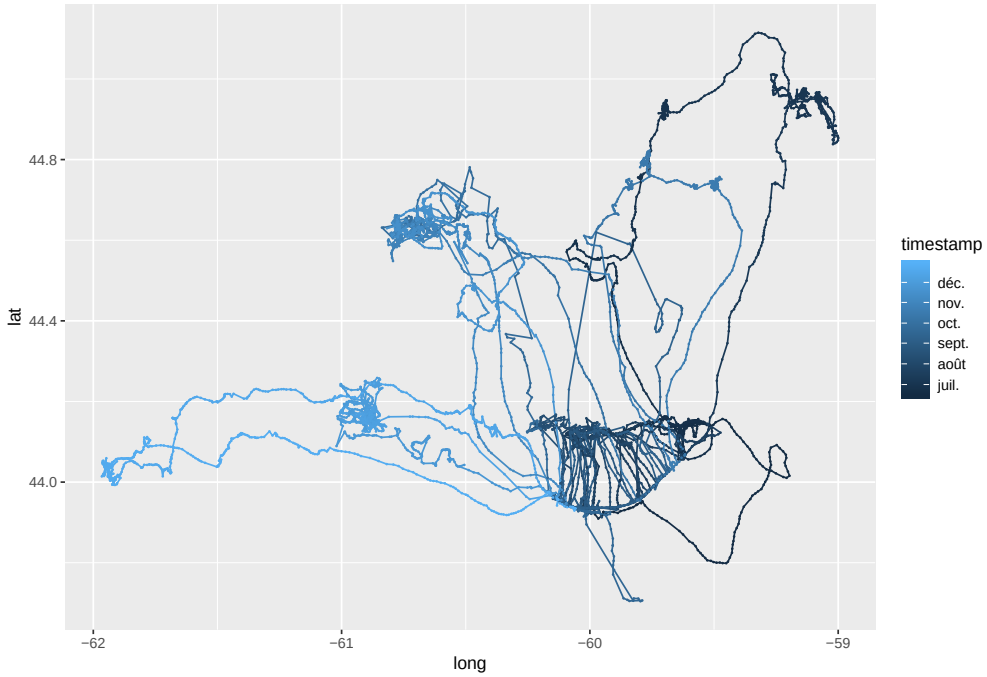


Figure 3.3: Path of seal F757 from [Bak+15]

to 1600 seconds among individuals. Mild irregularity in sampling process is often ignored but, in many situations, the statistical analysis has to cope with this sampling irregularity.

3.1.3 Movement on Earth

The tracked individuals move on the earth, their positions are recorded through geographic coordinates, with sometimes additional information on the altitude or the depth when it is relevant. Dealing with geographic positioning might be difficult as classical metrics (like euclidean distance for example) are not relevant anymore but in most situation, it is possible to define a specific local projection which does not affect the geometric property of the movement greatly. It is more difficult for large migratory species or for individuals traveling around the geographical poles, where no accurate projection can be found.

3.1.4 From movement to movement model

Tracking data provide observations of the movement of an individual. As illustrated in Figure 3.4, those observations are the combination of

- the actual path of the individual, denoted by $\boldsymbol{X} = (X_s, s \geq 0)$,
- some sampling process $\mathscr{T}$, i.e a sequence of increasing times at which the position of the tracked individual is registered,
- some potential geolocation errors.

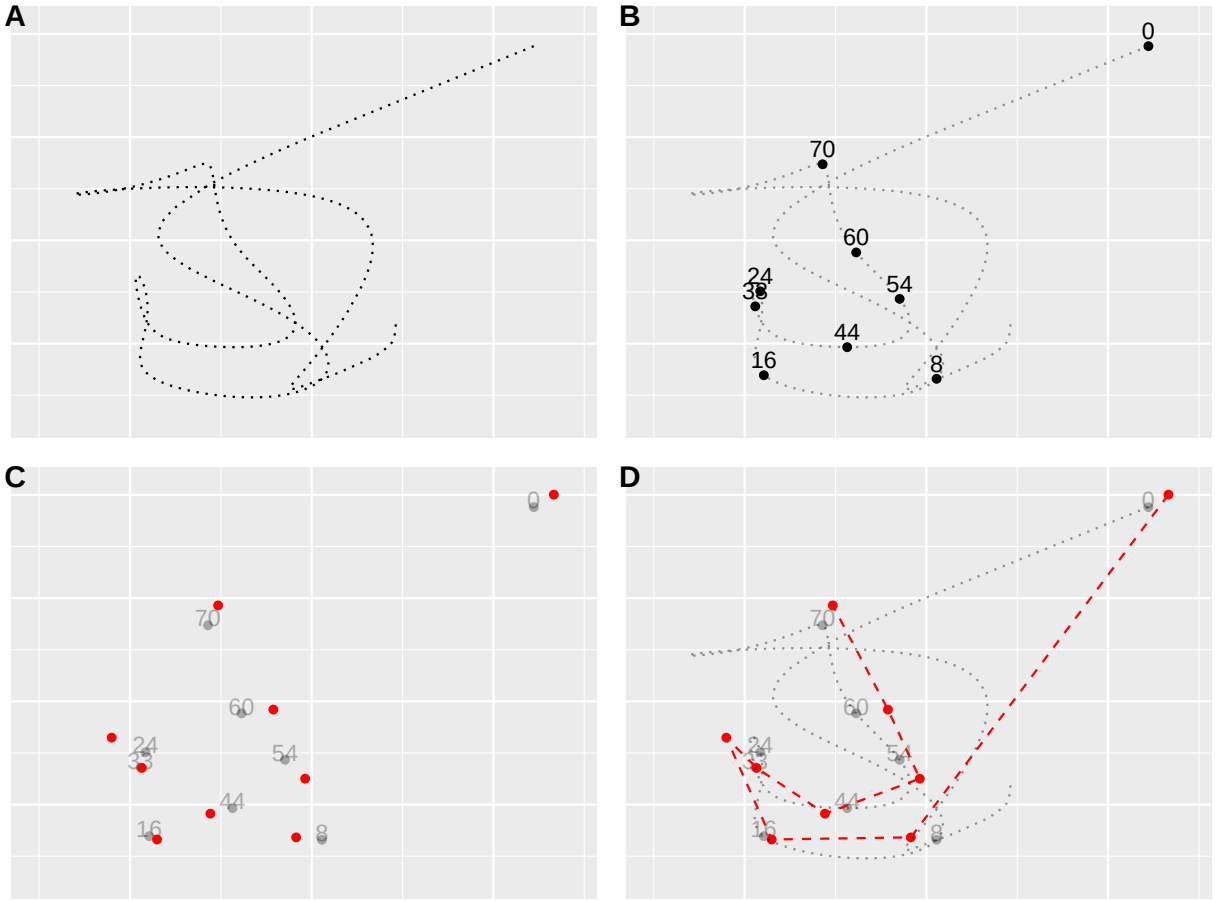


Figure 3.4: Figure A shows a realization of the random process $\boldsymbol{X}$. The sequence of sampled relocations $\boldsymbol{X}_{0:n}$ is given in the B plot (the numerical values corresponding to the realization of the sampled times). Figure C presents the recorded relocations at the sampling times, including measurement errors in red. Figure D compares the actual path with the classical linear interpolation of the registered path.

A realization of the sampling process $\mathscr{T}$ will be denoted by the sequence $(t_i)_{1\leq i\leq n}$. The individual path observed at those times will be denoted by $\boldsymbol{X}_{0:n} = (X_0, \ldots, X_n)$ where X_i stands for the actual position (i.e. considered as a point in $\mathbb{R}^2$) at time t_i, while the sequence $\boldsymbol{Y}_{0:n} = (Y_0, \ldots, Y_n)$ will stand for the corresponding sequence of recorded positions. If the positions are recorded with no error, $X_i = Y_i$ for all $i = 0, \ldots, n$.

Movement data are the result of a continuous time, continuous space process observed through a sampling process and therefore form a time series of georeferenced data. Therefore statistical methods for movement analysis borrow from the method of spatial point pattern statistics, from time series methods, or from continuous time stochastic processes.

3.2 Some fundamental movement ecology concepts

3.2.1 Utilization distribution

A crucial concept in animal ecology is the utilization distribution, *the probability density function that gives the probability of finding an animal at a particular location* [And82]. Proposing reliable method to predict utilization distribution is of major importance for wildlife management, for example to define Marine Protected areas or land use planning regulations.

Formally, as$X_t \in \mathbb{R}^d$ denotes the location of an animal in d-dimensional space at time $t \geq 0$, and $\pi : \mathbb{R}^d \to \mathbb{R}$ its utilization distribution [Wor89]. The utilization distribution is the probability density function π which satisfies

$$\mathbf{P}\left\{X_t \in A\right\} = \int_A \pi(z)\,\mathrm{d}z, \qquad \text{(Eq 3.2.1)}$$

for any area $A \subset \mathbb{R}^d$. Assuming that the individuals are in a steady state, π would correspond to the stationary distribution of the movement process $\boldsymbol{X}$. Thus the concept of use distribution in itself assumes the existence of a stationary regime. This very strong assumption is a cornerstone of movement ecology even though it is quite questionable since individuals evolve in a changing environment.

3.2.2 Home range

Burt [Bur43] defined the home range as *that area traversed by an individual in its normal activities of food gathering, mating, and caring for young. Occasional sallies outside the area, perhaps exploratory in nature, should not be considered part of the home range.*

There is no formal definition of home range and, as there exist different methods for estimating it, there also exist different mathematical definitions. Although the formal definition of the home range is a matter of debat [Kie+10; PM12], here I adopt the definition based on the utilization distribution. The home range is defined as the smallest connected space bordered with iso-probability density lines which contains at least 95% of the utilization distribution.

3.2.3 Resource selection function

A Resource Selection Function (RSF) is originally defined as *any function that is proportional to the probability of use by an organism* [MMT93]. A resource selection function is any function that links the characteristics of a spatial unit to its use. If a spatial unit x is fully characterized by J environmental variables $(c_1(x), \dots, c_J(x))$, a classical form for this function is

$$\pi(x|\beta) = \frac{\exp\left(\sum_{j=1}^{J} \beta_j c_j(x)\right)}{\int_\Omega \exp\left(\sum_{j=1}^{J} \beta_j c_j(z)\right) \mathrm{d}z},$$

where $(\beta_j)_{j=1,\dots,J}$ capture the effect of the different environmental covariates.

3.3 Movement models

Spatial statistics and more specifically point processes approaches have been widely used to understand how individuals use space. Kernel density estimation (KDE) is a non parametric approach [Fle+15; FC17] used to identify the home range of an individual. This method considers the spatial repartition of the sequence $\boldsymbol{X}_{0:n}$ as the realization of point process. The link between movement and point process is rarely explicitly considered except to account for some spatial dependence between points. I won't develop those approaches in this document but the interested reader could refer to [Hoo+17, Chapter 4].

Methods based on time series analysis directly model the sequence $\boldsymbol{X}_{0:n}$, i.e. without distinguishing between the actual movement process on the one hand and the sampling process on the other hand while continuous time models propose to use continuous time stochastic processes to model the actual movement of an individual $\boldsymbol{X}$. I intend now to describe briefly[3] the major advances that have been made in the recent years, using dynamic model for movement analysis and how my work fits into it. This presentation is structured as follows. First, Section 3.1.4 presents the classical movement models and how those models have evolved in the last decade. The next section presents two solutions to account for internal states. Where relevant, I will present how the environment is taken into account.

3.3.1 Discrete time models

As clearly stated in McClintock et al. [McC+14], although movement occurs in continuous time, it is often observed at almost fixed discrete-time intervals and it might be more intuitive to interpret movement in discrete time. This approach has been very popular and widely used from the start of the '90s. Mostly, rather than analyzing absolute position, different metrics are derived from the sequence $\boldsymbol{Y}_{0:n}$, such as

- the step length sequence, $\boldsymbol{L}_{1:n} = (L_1, \dots, L_n)$, $L_i = \|Y_i - Y_{i-1}\|$ being the distance between two successive recorded locations,

[3] A more extensive presentation of the different methods can be found in Hooten et al. [Hoo+17].

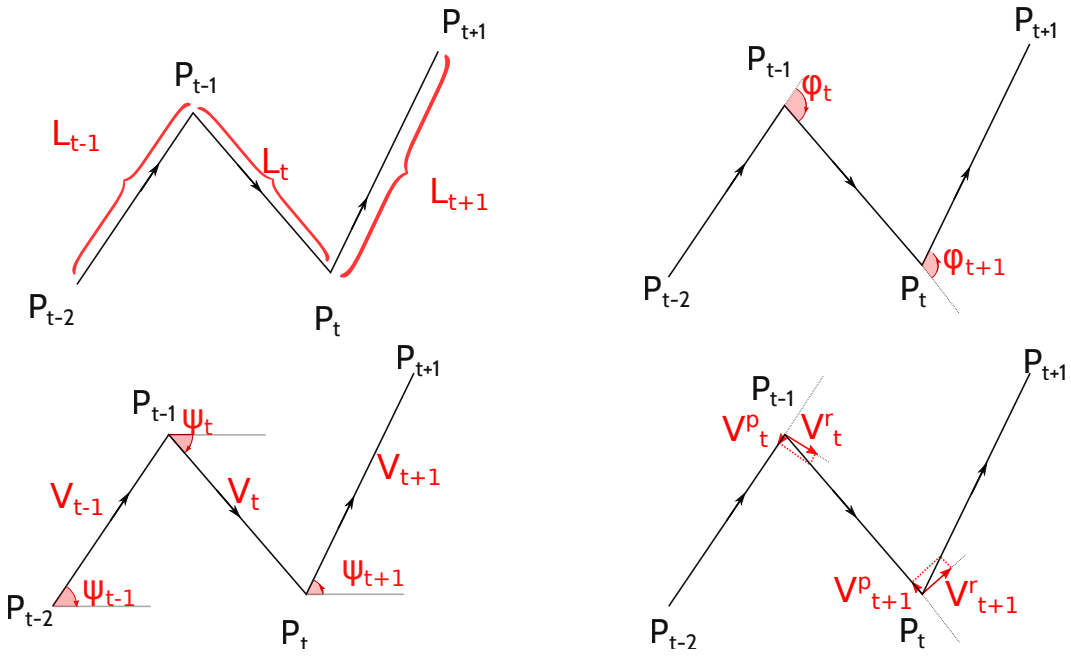


Figure 3.5: Representation of the classical metrics associated with movement decomposition in discrete time movement model.

- the absolute compass orientation ($\boldsymbol{\psi}$),
- or the turning angle sequence $\boldsymbol{\phi}$, ϕ_i being the difference between successive compass orientation, see for example Vermard et al. [Ver+10] and Bez et al. [Bez+11],
- or orthogonal components of persistence velocity $\boldsymbol{V^p}$and rotational velocity $\boldsymbol{V^r}$as in Gurarie, Andrews, and Laidre [GAL09] and Gloaguen et al. [**Glo+15**]

Those different metrics are illustrated on Figure 3.5.

As mentioned before, these discrete time models are said to be easier to interpret [McC+14] but this ease of interpretation might be misleading. First the step length or equivalently the average speed between two successive positions might be a very poor proxy of the actual movement as illustrated in figure 3.4 and, when possible the sampling frequency has to be chosen carefully. Since movement and sampling scheme are modeled altogether, the estimated parameters are tied to this temporal scale. Therefore, the comparison of such parameters between different studies (same species at different sites for example) might be difficult if even possible. Finally, discrete time models assume that the observations are made over regular time interval and this assumption is crucial since the sampling process in integrated in the model. This assumption might vary from slightly to highly unrealistic depending on the context. Schlägel and Lewis [SL16] explored the model's capacities to compensate for varying temporal discretization and defined a notion of robustness. They showed that few movement models have this robustness property. In practice, some pre-processing is performed to obtain equal time intervals, for instance by applying some linear

or more complex interpolation [JFM05; **Glo+15**]. This regularization step is popular and available in the classical package `adehabitatLT`. Some authors have advocated for its interest [BB88] however it might also considerably modify the actual animal movement and since this process is not accounted for in the model, the uncertainty associated with it remains unknown.

As a conclusion, although discrete movement models are widely used and appear to have simpler mathematical and conceptual formalization, irregular sampling times and comparison between different studies will benefit of the development of continuous time partially observed movement models.

3.3.2 Continuous time models

Once we agree on the need for continuous time models, since movement data are by nature spatio temporal data, we still have to address the question of space. Few papers propose to discretize the space. Trying to link covariates and movement, Hooten et al. [Hoo+10] and Hanks, Hooten, Alldredge, et al. [HHA+15] argue that covariates might be only available at some grid resolution (classical for rasterized data) and therefore advocate for a discrete space continuous time movement model and propose to use a discrete space continuous time Markov process. This process jumps from cell i to some adjacent cell j with intensity $\lambda_{ij} = \exp\left\{\boldsymbol{z}'_{ij}\boldsymbol{\beta}\right\}$, where $\boldsymbol{z}_{ij}$ stands for the vector of cell properties (the covariates, available at the cells scale) and $\boldsymbol{\beta}$ is vector of parameters which describe the effect of the different covariates. Using a data augmentation trick, the authors propose to represent the continuous time Markov chain as a Generalized Linear model and therefore propose a highly efficient algorithm to estimate this model. Since the discretization is driven by the availability of the covariates, this might be more difficult (or require very fine discretization) if covariates are available at different cells scale. The grid resolution thus becomes a parameter to be tuned.

Since animals move continuously in time and space, continuous time and continuous state models appear to be the most natural modeling framework. This framework has received a significant amount of attention over the past two decades and most of my research work in the field of movement ecology falls within this framework.

As early as 1952, Wilkinson [Wil52] investigated the potential of **Brownian motion** as a type of movement to explain the homing success of displaced bird. As explained in section 1.2.1, BM is central in the theory of stochastic processes and has been widely studied [RY13]. BM motion has zero mean and independent Gaussian increments. From a modeling point of view, assuming that the movement is accurately described by a BM implies that there is no trend in the movement and no memory. Furthermore, as a direct consequence the step length should verify $L_i^2 = \left\|W_{t_i} - W_{t_{i-1}}\right\|^2 = \sum_{p=1}^{2}(W_{t_i}^p - W_{t_{i-1}}^p)^2$, and the normalized step length L_i^2/Δ_i should exhibit a χ^2 distribution. In Figure 3.3.2, I compare the histogram of the normalized step lengths form the the brown bear data set available in the `adehabitatLT` [Cal06] with their expected distribution. With few surprise the two distributions don't match very well. This model is almost never explicitly used as a movement model, however it might be implicitly used in different approaches. For example, one classical method to infer Utilization Distribution from movement data, has

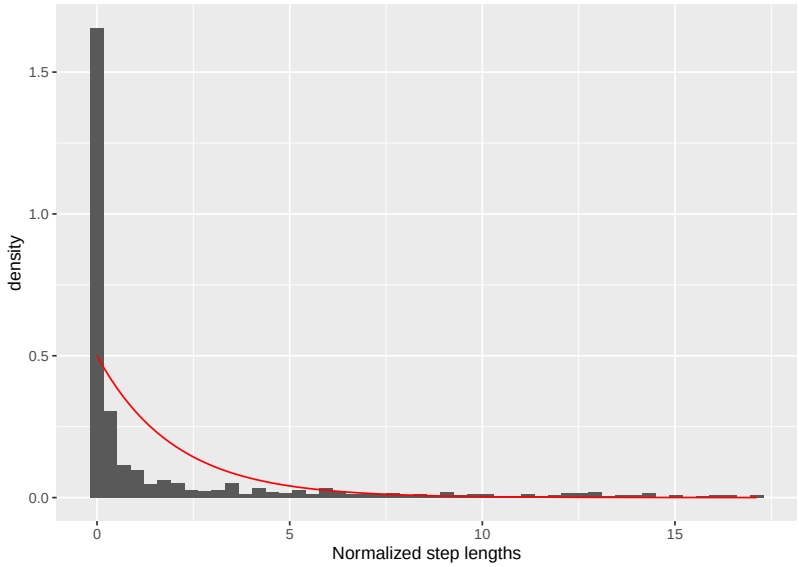


Figure 3.6: Histogram of normalized step length from a female brown bear monitored using GPS collars during May 2004 in Sweden, available in the `adehabitatLT` package. The red line is the density of a χ^2 distribution with 2 degrees of freedom.

been proposed by Horne et al. [Hor+07] and is based on a **Brownian bridge** approach and consequently assumes implicitly that the animal movement is at least locally similar to a BM.

As stated in introduction , **drifted Brownian motion** $\boldsymbol{W^\mu}$ is a natural extension of Brownian motion. which corresponds to a Brownian motion plus a deterministic constant drift and is defined as

$$\boldsymbol{W^\mu} = (W_t^\mu, t \geq 0), \qquad \text{with } W^\mu = \mu t + W_t.$$

Some extensions of the Brownian bridge approach for UD estimation have been proposed by [Ben11] which rely on the assumption that movement is a succession of Brownian Motion with drift. To understand how it improves the classical Brownian bridge it is interesting to represent the drifted Brownian motion as the solution to the following Stochastic Differential Equation (SDE),

$$\mathrm{d}\boldsymbol{W^\mu} = \mu \, \mathrm{d}t + \mathrm{d}W_t.$$

This formulation using SDE emphasizes that the drift term is the deterministic part of the movement, while the diffusion represents the unexplained aspects of the movement. The drift term can be understood as the deterministic component of the individual velocity. Identifying some mechanisms in the movement, within the SDE framework, would therefore consists in identifying and estimating a relevant form for the drift term.

Most animals belong to a given area (the home range), and some are linked to a colony. A natural modification of the Brownian motion with drift would consist in adding an attraction point in the drift, and define the corresponding **Ornstein Uhlenbeck** process. A a 2-

dimensional OU is the solution to:

$$\mathrm{d}U_t = -B(U_t - \mu)\,\mathrm{d}t + \Sigma\ \mathrm{d}W_t, \tag{Eq 3.3.2}$$

where B and Σ are two matrices in $\mathbb{R}^2 \times \mathbb{R}^2$.

The drift term depends on the difference between the actual position U_t and the attraction point. If the matrix B has only positive eigenvalues, this model represents a central behavior and admits a stationary distribution. The farther away the animal is from this center of attraction, the greater the force that recalls it back to that point. This movement model is sometimes named Ornstein-Uhlenbeck position model [Pat+17]. It has the desirable property to converge to its unique stationary distribution. This equilibrium distribution is a Gaussian distribution with mean μ and variance Σ_{stat} given by:

$$vec(\Sigma_{stat}) = (B \oplus B)^{-1} vec(\Sigma\Sigma^\intercal),$$

where $\oplus$ stands for the Kronecker sum as defined in Definition 2, p. 102.

The **Integrated Ornstein Uhlenbeck process** also named Ornstein Uhlenbeck velocity model has been introduced in movement ecology by [Joh+08] and has been recently proved to be a robust and accurate movement model [Gur+17]. In this model, the velocity is solution to equation Eq 3.3.2, and the location of the individual can be found by integrating the velocity over time:

$$X_t = \int_0^t U_t\,\mathrm{d}t,$$

with (U_t) solution to Eq 3.3.2.

Those simple movement models provide a natural representation of the evolution of the velocity (especially when used as emission distribution in a HMM context as in Gurarie et al. [Gur+17]) but have also been popularized because the estimation of such models is straightforward. Let's consider an observed sequence of relocations, $(X_0, \ldots, X_n)$, the log likelihood of this sequence is defined by

$$\ell(\theta; \boldsymbol{X}_{0:n}) = \sum_{i=0}^{n-1} \log\left\{q_{\Delta_i}(X_i, X_{i+1})\right\},$$

and implies the transition q_{Δ_i} from location X_i to location X_{i+1} during time $\Delta_i = t_{i+1} - t_i$. When X is defined as a solution to a SDE, this transition is generally unknown except in a few specific cases that include the models mentioned above, which fail to represent the actual attractiveness of the environment except the effect of some central location attractivity. Brillinger et al. [Bri+02] proposed a SDE framework which reflects the attractiveness of the environment.

3.3.3 Stochastic Differential Equation for movement model

As mentioned above, the drift term of a SDE counts for the determinisms responsible for the movement. In [BSL+08], the authors propose a potential based model:

$$\mathrm{d}X_t = -\nabla H(X_t)\,\mathrm{d}t + \gamma\,\mathrm{d}W_t, \quad X_0 = x_0 \tag{Eq 3.3.3}$$

where H is a potential function which represents the attractiveness of the environment.

This potential function based model is linked to the utilization distribution. If $\int_{\mathbb{R}^2} \exp -H(x)$ is finite, then the solution to the SDE defined above admits a stationary distribution π defined by:

$$\pi(x) = \frac{\exp\{-2\gamma^{-2}H(x)\}}{\int_{\mathbb{R}^2} \exp\{-2\gamma^{-2}H(u)\}\,\mathrm{d}u}.$$

The Ornstein Uhlenbeck process $\boldsymbol{U}$ defined in Eq 3.3.2 is a specific case of this potential based SDE where the potential H equals $\frac{1}{2}(\mu - X)^\intercal B(\mu - X)$ and admist a Normal distribution as utilization distribution. This modeling approach encompasses a large number of other processes and allows more flexibility for movement modeling. Since the transitions of a potential based SDE are unknown in general, the estimation problem remains. As it is classically done in application of SDE in finance, Brillinger et al. [Bri+02] use the Euler–Maruyama method. This method produces biased estimation but this bias vanishes when the maximal time between successive observations, $\Delta^* = \max_i |\Delta_i|$, tends to 0 ([KLS12]) which is the case in most financial applications (especially in the context high frequency trading). However, in movement ecology, the discrepancy between two successive observations is highly variable as illustrated in the dataset in Table 3.1.

In collaboration with Pierre Gloaguen and Sylvain Le Corff, we examined, in [**GEL182**], the question of the best estimation method in the context of movement ecology and propose a new flexible model, the GaP model, to describe the position process of an individual. $\boldsymbol{X}$ is assumed to be the solution to the following time homogeneous SDE:

$$X_0 = x_0, \quad \mathrm{d}X_t = -\nabla H_\eta(X_t)\,\mathrm{d}t + \gamma\,\mathrm{d}W_t, \tag{Eq 3.3.4}$$

where $\gamma \in \mathbb{R}_+^\star$ is an unknown scalar diffusion parameter and we assume that the potential H_η is a mixture function

$$H_\eta(x) := \sum_{i=1}^{K} \pi_k \phi_k^\eta(x) \text{ with } \phi_k^\eta(x) := \exp -\frac{1}{2}(x - \mu_k)^\intercal C_k (x - \mu_k), \tag{Eq 3.3.5}$$

where

- K is the number of components of the mixture;
- $\pi_k \in \mathbb{R}^+$ is the relative weight of the k-th component with $\sum_{k=1}^K \pi_k = 1$;
- $\mu_k \in \mathbb{R}^2$ is the center of the k-th component;
- $C_k \in \mathscr{S}_2^+$ is the information matrix of the k-th component, where $\mathscr{S}_2^+$ is the set of 2×2 symmetric positive definite matrices.

Because we assumed that the observation error is small enough to be neglected, the observed sequence $\boldsymbol{Y}_{0:n} = \boldsymbol{X}_{0:n}$. We compared the Euler method with two other pseudolikelihood procedures: the Ozaki method, [Oza92] which proposes to improve the Euler scheme by a local linearization of the drift term, and the Kessler method [Kes97] which generalizes

the Euler–Maruyama method by using a normal transition density whose mean and variance are chosen equal to the actual mean and variance. We also compared with an exact Monte Carlo Expectation Maximization algorithm (EAMCEM) approach based on exact simulation of diffusions proposed by Beskos et al. [Bes+06].

The quality of the estimation is addressed through simulation studies with varying sampling scheme of the movement stochastic process and we proved that the Euler method performs worse than all other procedures for all sampling schemes, but the difference vanishes Δ tends to 0. The application of this model on two French fishing vessels movement produce slightly different estimated potential maps, especially when using the Euler method as illustrated in Figure 3.7. The Ozaki method showed results similar to those of the exact algorithm based Monte Carlo EM approach.

Using Vessel Monitoring System[4] (VMS) data, the GaP model has been used by Pierre Gloaguen in his PhD thesis [Glo15] to produce a subjective map of Cuttlefish relative abundance. However this map was not correlated with the scientific abundance map produced using scientific campaign abundance monitoring.

The parametric form of the GaP model has been chosen to provide a flexible movement model which permit to compare different estimations algorithms, therefore it had to verify the conditions required to apply the exact simulation algorithm EA1 proposed by Beskos et al. [Bes+06]. By this choice, the existence of a utilization distribution was abandoned. However, the alternative estimation methods do not impose this restriction. A flexible model which admits a stationary distribution is then given by the SDE based on the following potential drift function:

$$H(x) = H_\eta(x) + (x-\nu)^\intercal \Sigma (x-\nu),$$

where $\nu \in \mathbb{R}^2$ is parameter which reflects some central place behavior, and $\Sigma \in \mathscr{S}_2^+$. H is now an integrable potential function and is associated to a utilization distribution.

3.3.4 Partially observed SDE

As mentioned in section 3.1.1, depending on the technology, the accuracy of the recorded location can be very questionable, especially while working with ARGOS data. The model proposed in the previous section does not account for observation errors. In practice, the data are mostly pre-processed using different procedures such as the State Space methods [Fre+08; Pat+10] or a continuous time correlated random walk (the discrete counterpart of the Ornstein Uhlenbeck velocity model presented in the 3.3.2 as proposed by Johnson et al. [Joh+08] and Johnson and London [JL18]). After this pre-processing step, the data are analyzed with the identified relevant method. Doing so, the smoothing induced by the pre-processing step is ignored which can be misleading.

In [**GEL181**], in collaboration with the same colleagues, I initiated a first step towards an integrated approach to estimate the parameters of a movement model solution to a SDE but observed with error. We defined a general algorithm to conduct inference for hidden Markov

[4] VMS is a *satellite-based monitoring system which provides location data to the fisheries authorities at (more or less) regular intervals (mostly every hour*

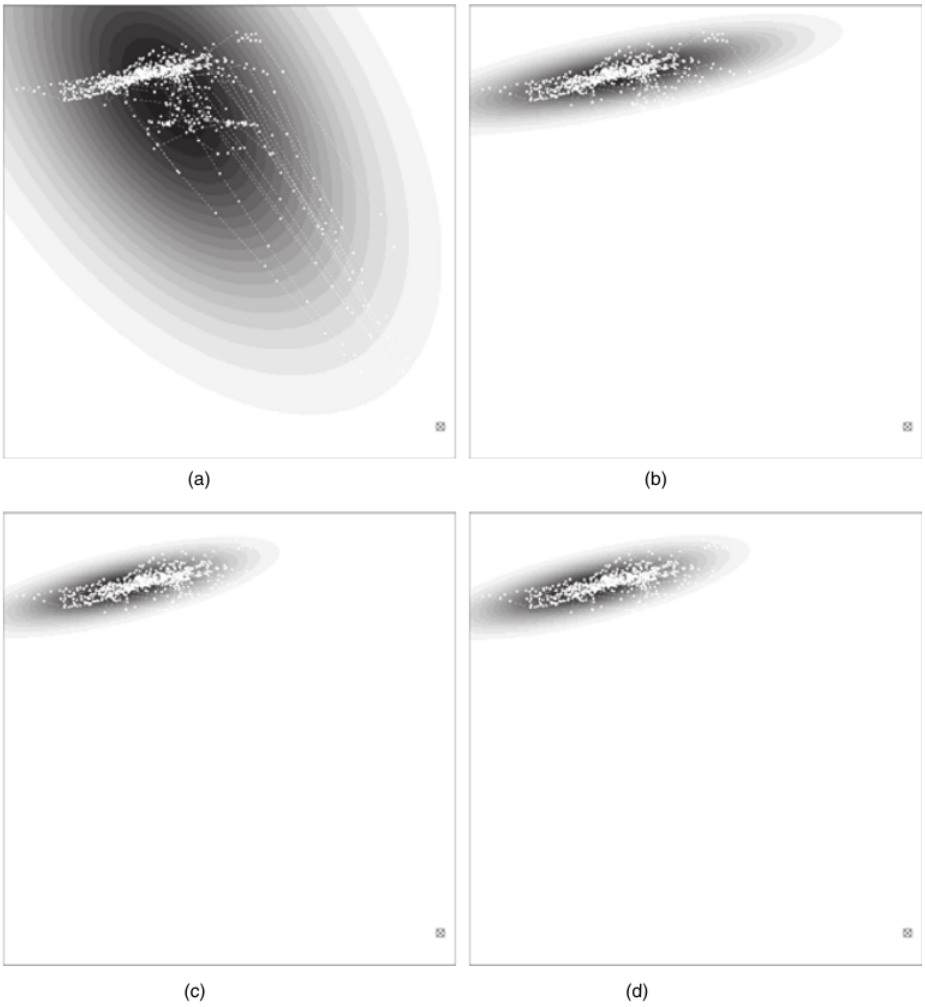


Figure 3.7: Potential map estimated from 2 French vessels tracks using four different estimation methods (x, departure harbor) (the darker a zone is, the more attractive it is for the given vessel; observed points are plotted in white to see the superposition between maps and trajectories): (a) Euler; (b) Kessler; (c) Ozaki; (d) EAMCEM. Figure extracted from [GEL182].

models (HMMs) whose hidden state is a solution to a stochastic differential equation (SDE). These models are referred to as partially observed diffusion (POD) processes in [OS+11]. As mentioned before, the Expectation Maximization (EM) algorithm introduced by [DLR77] is a classical algorithm to infer parameters implying latent variables and is widely popular in the HMMs context [Rab89; CMR06]. This algorithm relies on the conditional distribution of the hidden variables given the data.

In the HMMs context, it is useful to define the smoothing and filtering distributions. For all $0 \leq k \leq k' \leq n$, the joint smoothing distributions of the hidden states are defined, for all measurable function h on $(\mathbb{R}^d)^{k'-k+1}$, by:

$$\varphi_{k:k'|n}[h] = \mathbf{E}\left\{h\left(X_k, \ldots, X_{k'}\right) | \boldsymbol{Y}_{0:n}\right\},$$

and $\varphi_k := \varphi_{k:k|k}$ denotes the filtering distributions.

For HMMs with the snapshot property (random variables $Y_{0:n}$ given the hidden states $\boldsymbol{X}_{0:n}$ are conditionally independent), the pivotal quantity Q is expressed as an additive functional of the hidden states given all the observations up to time n:

$$Q(\theta, \theta^{(i-1)}) = \varphi_{0|n}[h_1] + \sum_{k=0}^{n-1} \varphi_{k:k+1|n}[h_2] + \sum_{k=1}^{n} \varphi_{k|n}[h_3],$$

where $h_1(X_0) := \ln p_\theta(X_0)$, $h_2(X_k, X_{k+1}) := \ln p_\theta(X_k|X_{k-1})$ and $h_3(X_k) := p_\theta(Y_k|X_k))$.

Sequential Monte Carlo (SMC) methods are popular algorithms to approximate filtering and smoothing distributions with random particles associated with importance weights [GSS93]. In the specific case of HMMs, approximations of the smoothing distributions may be obtained using the forward filtering backward smoothing algorithm (FFBS) and the forward filtering backward simulation algorithm (FFBSi) developed respectively in [DGA00] and [GDW04]. Recently, Olsson, Westerborn, et al. [OW+17] proposed a new SMC algorithm, the particle-based rapid incremental smoother (PaRIS), to approximate on-the-fly (i.e., using the observations as they are received) smoothed expectations of additive functionals.

Unfortunately, these methods cannot be applied directly to POD processes since some elementary quantities, such as transition densities of the hidden states, are not available explicitly. We proposed in Gloaguen, Etienne, and Le Corff [**GEL181**], the GRand PaRis algorithm, an extension of the PaRIS algorithm where the exact transition densities are replaced by unbiased estimators using generalized Poisson estimators (GPE) proposed by Fearnhead, Papaspiliopoulos, and Roberts [FPR08] and we proved that the acceptance rejection mechanism introduced by Douc et al. [Dou+11] ensuring the linear complexity of the procedure is still correct when the transition densities are replaced by those unbiased estimates. We proposed two simple applications of this algorithm for one dimensional POD.

However the PaRIS, and the GRand PaRIS algorithm as well require that the unbiased estimate of the transition is almost surely positive and bounded to perform the crucial acceptance rejection step. In the context of SDE, this assumption is very restrictive and narrows the possible models to the class of diffusion satisfying the Exact algorithm conditions of Beskos et al. [Bes+06], for which GPE leads eligible unbiased estimators. In a recent work,

in collaboration with Pierre Gloaguen, Sylvain Le Corff and Jimmy Olsson, we proposed to replace the backward acceptance-rejection step by an importance sampling estimate which leads to a smoothing algorithm that only requires almost surely positive estimator of the transition densities, which can be obtained for a wide range of diffusion processes using the parametrix estimators of Andersson, Kohatsu-Higa, et al. [AK+17] and Fearnhead et al. [Fea+17]. In Etienne et al. [**Eti+21**], we show that the IS-PaRIS algorithm is faster than the GRand-PaRIS algorithm on a sine model. We also illustrate that it can be used to estimate (through an EM algorithm) a bivariate SDE observed with noise. We chose a Lokta-Volterra system observed with noise as an example, the observation being the number of hares and lynx trapped in Canada during the first 20 years of the 20th century (available in [OB71]).

This algorithm is very promising and solves an important statistical problem because it extents the class of POD process which can be estimated without discretization bias and decreases the computational cost compared to other POD smoother. However it is currently not fast enough to be used with actual movement ecology data (and so are the other unbiased POD smoothers) and there is still some place for improvement to propose unbiased and fast algorithm for POD.

3.3.5 Accounting for environment

The GaP model presented in section 3.3.3 is a flexible model to estimate the potential drift function. This potential is linked to the utilization distribution: As many other popular approach which study habitat preferences and movement, the density function is estimated thanks to telemetry data, and this density function can be related to environmental covariates using regression techniques [Mil+06; Lon+09; Zha+14] in a second step. As detailed in section 3.2.3, this regression function is often defined as a resource selection function [Man+02]. It is based on the idea that, knowing the habitat composition of a spatial unit, we can predict its long-term utilization. However, this two step procedure is not satisfying because it ignores the uncertainty on the estimated utilization distribution.

It is natural to think of the utilization distribution as a consequence of the movement, which itself depends on the environment, such that short-term movement decisions give rise to long-term space use. This idea motivates the development of more mechanistic approaches that link the animal's movement to its environment, and, ultimately, a mechanistic movement model with an explicit steady-state distribution, representing the utilization distribution. The step selection functions model the local movement (a step) as a combination of a movement kernel and a habitat selection function [TCB14]. The habitat preferences are assessed in comparison with other potential movements, described by a movement kernel. However defining what the other potential movements remains a major concern [Mat03; Lel+13; Nor+13] and up to recently only very simple movement model has been used to define potential movement. Recently, Avgar et al. [Avg+16] proposed an integrated approach and assumed that animal movement can be represented by a separable model, more precisely the product of a discrete time movement kernel and a a selection kernel and proposed to estimate simultaneously these two kernels in an integrated Step Selection Analysis

(iSSA). Hanks, Hooten, Alldredge, et al. [HHA+15] proposed a continuous-time discrete-space model to link movement to environmental drivers. In their framework, the movement is considered as a continuous-time Markov process on a discrete grid of spatial cells. The spatial grid is usually chosen as the grid on which the spatial covariates are measured, and the observed locations are binned in the cells. Wilson, Hanks, and Johnson [WHJ18] argued that the limiting distribution of that movement model can be interpreted as the utilization distribution of the animal, and proposed a method to estimate it on a discrete grid. These two approaches described movement on a discrete spatial grid, and their formulation is therefore tied to a particular space discretization. Recently, [MBM19] proposed a step selection model, formulated in terms of an explicit utilization distribution. Their approach described individual movement as a Markov chain in continuous space, whose stationary distribution is the utilization distribution. In particular, they suggested that Markov chain Monte Carlo (MCMC) algorithms, which are used to construct Markov chains with a target stationary distribution, can be viewed as movement models. These proposals are implemented in computationally intensive approaches.

In collaboration with Théo Michelot and Pierre Gloaguen [**Mic+19**], we brought together the ideas of [Bri+02] and [MBM19] and propose a new model based on the Langevin diffusion, which has also been used to construct an MCMC algorithm [RR98]. We assume that $\boldsymbol{X}$ is the solution of the following SDE:

$$\mathrm{d}X_t = \frac{\gamma^2}{2}\nabla \log \pi(X_t)\,\mathrm{d}t + \gamma\,\mathrm{d}W_t, \quad X_0 = x_0, \tag{Eq 3.3.6}$$

where $\gamma \in \mathbb{R}^+$ is a scaling parameter which might be thought as the natural speed of the individual. The solution to this SDE is a stochastic process which admits π as stationary distribution. Formulating the same idea from a movement ecologist prospective, assuming that the movement of an individual verifies Equation Eq 3.3.6, this individual admits π as utilization distribution function.

We link the utilization distribution π of the animal to spatial covariates with the standard parametric form of RSF:

$$\pi(x|\beta) = \frac{\exp\left(\sum_{j=1}^{J}\beta_j c_j(x)\right)}{\int_\Omega \exp\left(\sum_{j=1}^{J}\beta_j c_j(z)\right)\mathrm{d}z}, \tag{Eq 3.3.7}$$

where $c_j(x)$ is the value of the j-th covariate at location x, $\Omega \subset \mathbb{R}^d$ is the study region, and $\beta = (\beta_1, \ldots, \beta_J)'$ is a vector of unknown parameters. The denominator in the right-hand side of Equation (Eq 3.3.7) is a normalizing constant, and is necessary to ensure that $\pi(x|\beta)$ is a probability density function with respect to x.

Using Equations Eq 3.3.6 and Eq 3.3.7, we propose the following model:

$$\mathrm{d}X_t = \frac{\gamma^2}{2}\sum_{j=1}^{J}\beta_j \nabla c_j(X_t)\,\mathrm{d}t + \gamma\,\mathrm{d}W_t, \quad X_0 = x_0. \tag{Eq 3.3.8}$$

As this model belongs to the class of potential-based models, inference can be performed from movement data using different estimation methods for stochastic differential equations

(SDEs), such as pseudo-likelihood methods as presented in [GEL182]. However, thanks to the RSF classical form, the Euler approximation has a very simple form

$$X_{i+1}|\{X_i = x_i\} = x_i + \frac{\gamma^2\Delta_i}{2}\sum_{j=1}^{J}\beta_j\nabla c_j(x_i) + \sqrt{\Delta_i}\varepsilon_{i+1}, \quad \varepsilon_{i+1} \overset{ind}{\sim} N\left(0, \gamma^2 I_d\right), \quad \text{(Eq 3.3.9)}$$

where $\Delta_i := t_{i+1} - t_i$. This approximation leads to a standard linear model formulation for Y, the sequence of normalized increments of X:

$$Y = Z\nu + E, \quad \text{(Eq 3.3.10)}$$

where

- $Y_i = (X_{i+1} - X_i)/\sqrt{\Delta_i}$ is the two-dimensional normalized random increment of the process between t_i and t_{i+1},
- Z is the matrix defined by

$$Z := \frac{1}{2}(\mathbb{I}_2 \otimes D)\begin{pmatrix} \frac{\partial c_1(x_0)}{\partial z_1} & \frac{\partial c_2(x_0)}{\partial z_1} & \cdots & \frac{\partial c_J(x_0)}{\partial z_1} \\ \vdots & & & \vdots \\ \frac{\partial c_1(x_{n-1})}{\partial z_1} & \frac{\partial c_2(x_{n-1})}{\partial z_1} & \cdots & \frac{\partial c_J(x_{n-1})}{\partial z_1} \\ \frac{\partial c_1(x_0)}{\partial z_2} & \frac{\partial c_2(x_0)}{\partial z_2} & \cdots & \frac{\partial c_J(x_0)}{\partial z_2} \\ \vdots & & & \vdots \\ \frac{\partial c_1(x_{n-1})}{\partial z_2} & \frac{\partial c_2(x_{n-1})}{\partial z_2} & \cdots & \frac{\partial c_J(x_{n-1})}{\partial z_2}, \end{pmatrix},$$

  with $D = (d_{kl})_{1\leq k,l\leq n}$, with $d_{kl} = \sqrt{\Delta_{k-1}}$ if $k = l$ and 0 otherwise and where $\partial/\partial z_i$ denotes the partial derivative with respect to the i-th spatial coordinate,
- $\nu = \gamma^2\beta$.

The estimators for ν and γ^2 and, as a consequence, the estimator $\hat{\beta}$ of the original parameters are derived from standard linear model theory and, as thus, the computation time using this Euler approximation is equivalent to the one for fitting a linear regression model, thus very fast for standard data set sizes. We also proposed to use classical linear diagnostic to check the residuals and identify regions where the model do not capture the movement well.

As highlighted in [GEL182] the validity of the Euler approximation is sometimes questionable in movement ecology application. We use the Metropolis-adjusted Langevin algorithm (MALA) ratio [RT+96] to assess the accuracy of the Euler approximation.

The proposed integrated movement model is implemented in the Rhabit package [EGM20]. This model has been applied to the Steller sea lion data set provided by Wilson, Hanks, and Johnson [WHJ18] and the corresponding UD is presented in Figure 3.8. It is worth noting that high densities are found in areas where sea lions are not observed. This is a consequence of the inclusion of environmental covariates, although no sea lions have been observed in this area, the habitat is equally favorable.

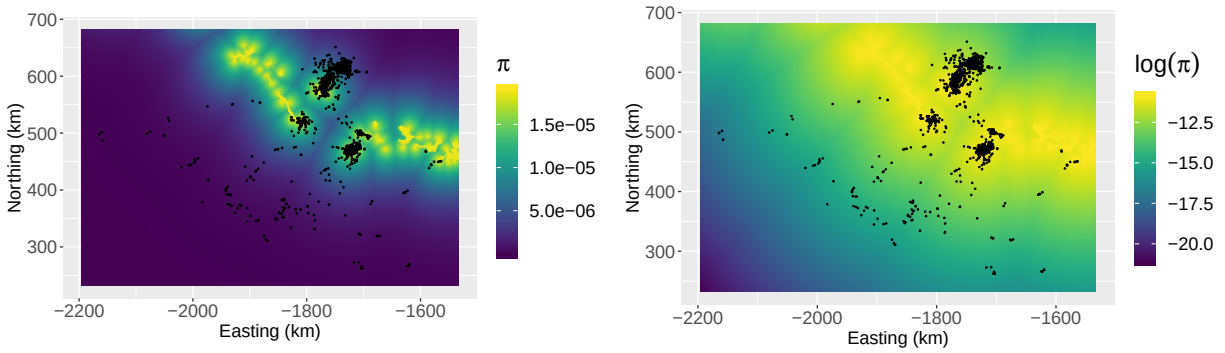


Figure 3.8: Estimated utilization distribution for Steller sea lion data set provided by [WHJ18] (left), and its logarithm, for comparison with the original publication (right). The black dots are the filtered sea lion locations.

3.4 Switching movement models

In the previous section, I described movement models and propose new models assuming that the movement model is unchanged during the whole observation period. But landscapes are spatially and temporally variable at various scales [Lev92], and animals are expected to adjust their movements to the characteristics of their local environment, so as to maximize the time spent in profitable (or safe) habitats and minimize time in adverse ones [Pyk78]. As illustrated by Nathan et al. [Nat+08] and recalled in Figure 3.2, movement is also influenced by internal factors. Consequently location time series, and/or the various series that can be derived from them to describe the movement behavior (e.g. turning angle, speed), are therefore expected to be only piecewise stationary.

Many studies have been developed to reveal behavior from movement data [Mor+04; GAL09] or reveal home range shift [Ben14; Cag+16]. In fisheries science the same ideas have been used to predict fishing states (fishing vs not fishing) based on VMS data [Ver+10; WB10] in order to produce a map of fishing effort. In all those different studies, the observed trajectory is understood as the realization of a succession of movement models governed by a succession of activities (foraging, feeding, resting by instance for movement ecology and fishing vs not fishing in fisheries science) or a succession of different central places in the Home range shift question[5]. The activities, represented by the process $\boldsymbol{S}$ in Figure 3.5, govern the movement of the individual.

Those activities are rarely observed except in very few cases like fishing vessels who can have on board observers. Although few studies have developed supervised learning approaches to predict behavior from movement data [LL12], they are quite uncommon in movement ecology. I should mention some application of those learning approaches for precision livestock farming [Rah+18] or connected equestrian equipment [Sch+20]. In

[5]In the remaining of this section, I will only refer to the term activities, but everything is applicable to the changes in HR.

such context, movement data are mostly coupled with accelerometer data. Thanks to the possibility to have on board observers, supervised learning approaches are more popular in fisheries science than in movement ecology [Mar+15; Rus+11; Joo+11]. In a context of a small scale fisheries, I have worked in collaboration with a PhD Student from Institut Halieutique et des Sciences Marines from the university of Toliara, Madagascar, to propose an efficient machine learning approach to quantify the spatial fishing effort [**Beh+21**].

Nonetheless, except in some very specific context, quite uncommon in movement ecology, we don't face a supervised learning problem but the question of identifying homogeneous portions of trajectories to be linked with behaviors remain. This question is classically addressed with one of the following approaches: Methods based on Hidden Markov Model and methods based on change point detection.

The DAG presented in Figure 3.5 can be formalized as follows. Let $\boldsymbol{S} = (S_t, t \geq 0)$ be a hidden process with finite space state $\mathscr{S} = \{s_1, \ldots, s_A\}$ which describes the activities of an individual. We denote by $J_1, \ldots, J_K$ the sequence of jumping times such that

$$J_k = \inf \left\{ t > J_{k-1}, S_t \neq S_{J_{k-1}} \right\}.$$

The movement model is assumed to stay unchanged on the interval $[J_k, J_{k+1}[$ and, as such, the distribution of the observed sequence on the same interval is characterized by a set of parameter depending on S_{J_k}

$$X_{i:j} | S_{J_k}, X_{t_{i-1}} \sim \mathscr{L}(\theta_{S_{J_k}}) \quad \text{for } J_k \leq t_i < \ldots < t_{i+j} < J_{k+1}.$$

Figure 3.9 shows an example of such model. The succession of 3 different movement regimes have been simulated with three different Ornstein Uhlenbeck processes to represent different central foraging places. Given the sequence of relocations, the goal is then to identify some homogeneous portion and estimate the parameters of the corresponding models.

3.4.1 Hidden Markov Model

As mention in section 1.2.2, the HMM approach is very popular to account for heterogeneity in time series and as so it has been widely used in many different applications [Mor+04; WB10; GAL09] as a classical approach for time series segmentation. In this context, $\boldsymbol{S}$ is assumed to be a discrete space Markov chain $\boldsymbol{S}_{0:n} = (S_0, \ldots, S_n)$ and, consequently the sequence of sojourn times $(J_k - J_{k-1}, k \geq 1)$ is a sequence of independent geometrically distributed random variables.

In most applications, the random variables $X_0, \ldots, X_n$[6] are assumed to be independent conditionally on the hidden states. In the movement ecology community, the inference is mainly conducted under a Bayesian framework and the predictions $\hat{\boldsymbol{S}}$ for the hidden sequence

[6] In many applications, the sequence $\boldsymbol{X}$ is not the sequence of relocations but some metrics derived from these relocations as presented in section 3.3.1 and illustrated in Figure 3.5. Thus $\boldsymbol{X}_{0:n}$ is the bivariate sequence of turning angles and speed for example, or persistent and rotational velocity.

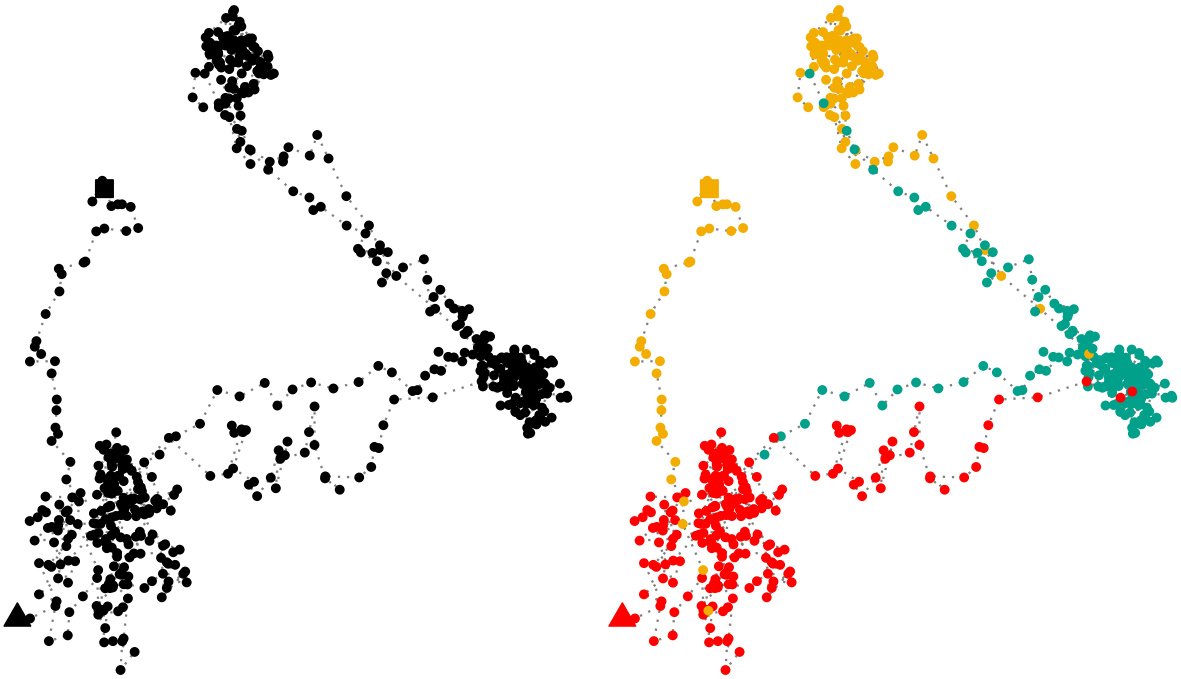


Figure 3.9: Simulated trajectory, resulting from a succession of three different Ornstein Uhlenbeck processes to represent variations in central foraging places. The triangle (resp. the square) indicates the start (resp. the end) of the sequence. On the right side, the relocations have been colorized according to the movement used to simulate.

are defined by the sequence of the individually most likely states i.e.

$$\hat{\boldsymbol{S}}_{0:n} = (\hat{S}_1, \ldots, \hat{S}_n), \quad \text{with } \hat{S}_i = \arg\max_{s_i} \mathbf{P}\left\{S_i = s_i | \boldsymbol{X}_{0:n}\right\}.$$

As part of Pierre Gloaguen's PhD thesis and in collaboration with Stéphanie Mahévas and Etienne Rivot, we analyzed French fishing vessels trips with an auto regressive process (AR) as emission distribution (the distribution of the observations) to describe the movement of fishing vessels [Glo15]. Contrary to most previous applications in the field, the estimation was conducted using the EM algorithm using the Viterbi algorithm, the state reconstruction was based on the mode of the smoothing distribution $\boldsymbol{S}$, i.e. the global most probable sequence given the observation and the estimated parameters:

$$\hat{\boldsymbol{S}}_{0:n} = \arg\max_{s_1,\ldots,s_n} p_{\hat{\theta}}\left\{S_1 = s_1, \ldots, S_n = s_n | \boldsymbol{X}_{0:n}\right\}.$$

Few differences were found between the two state reconstruction methods.

However, the HMM approach has several drawbacks. First, as a discrete-time model, it suffers from the common weaknesses of these models mentioned in section 3.3.1: difficulty to compare studies with different sampling times, regularity of the sampling process. Second, the Markov property implies that the sojourn time in a given state follows a geometric distribution. Finally, the discrete formulation of the model implies that the behavioral switch occurs simultaneously with the observation. This latter assumption is highly questionable in certain contexts, for example in the case of marine mammals whose positions are only recorded when they surface.

Using fisheries data, with on board observers and with a high sampling rate compared to the sojourn times, we have examined the relevance of the different assumptions and how semi Markov processes could improve the fit of such models [**Bez+21**]. We proved that the Markov models are robust to deviations from these assumptions.

To conclude this short section on the use of HMM for the segmentation of movement data, I would like to mention the package [MLP16] which provides a very flexible tools for segmenting movement data through an HMM approach, considering the movement as a succession of step length and turning angles. The model is fitted by numerically optimizing the likelihood with no help of the EM algorithm. Some environmental covariates can be included via a multinomial logit link in the transition matrix, the transition between states account for the environment.

3.4.2 Change point detection

An alternative to HMM and HSMM, which circumvents the difficulty of proposing relevant model for sojourn time is the change point detection method, which is a classical signal processing method.

Assuming that $\boldsymbol{S}$ is a constant piecewise function with $K-1$ changes at unknown times $\boldsymbol{\tau} = \tau_1, ..., \tau_{K-1}$ with the convention $\tau_0 = -1$ and $\tau_K = n$. $\boldsymbol{S}$ increases of one unit, at each change point. The value of the process $\boldsymbol{S}$ at time t indicates the identification number of the segment which contains t. This defines a partition of the data into K segments, segment k

being of length n_k. In any segment k, the sequence $\boldsymbol{X}_{\tau_{k-1}+1:\tau_k}$ is assumed to be a sequence of iid random variables with pdf parameterized by θ_k:

$$X_j \sim f_{\theta_k}(.), \quad \text{for } j \in \{\tau_{k-1}+1, \ldots, \tau_k\}, k = 1, \ldots, K, \tag{Eq 3.4.11}$$

Given the number of segments, the unknown parameters are the change points $\boldsymbol{\tau}$ and the set $\boldsymbol{\theta} = (\theta_k, k = 1, \ldots, K)$. They are estimated by maximizing the log-likelihood.

$$(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\theta}}) := \arg\max_{(\boldsymbol{\tau}, \boldsymbol{\theta})} \sum_{k=1}^{K} \sum_{t=\tau_{k-1}+1}^{\tau_K} \log f_{\theta_k}(x_k). \tag{Eq 3.4.12}$$

The brute-force algorithm consisting in optimizing the log likelihood over the $\binom{K-1}{n-1}$ possible values for $\boldsymbol{\tau}$ is numerically intractable. The problem is solved by the dynamic programming (DP) algorithm. An extensive presentation of this algorithm and a review of the most recent results can be found in [Leb18]. Figure 3.10 provides an illustration of such procedure for univariate signal.

Several methods exist to fix the optimal number of segments $K^\star$ based on model selection criterion like the modified BIC criterion proposed by Zhang and Siegmund [ZS07] or the largest slope change as proposed by Lavielle [Lav05].

A hierarchical extension of this model, named segmentation/clustering model has been proposed by Picard et al. [Pic+05] and applied to aCGH profiles (presented in the section 2) to identify sub-segment and decide whether or not a sub-segment should be classified as altered. More generally this approach assigns each segment to a cluster m. $\boldsymbol{S}_{0:n}$ takes only values in $\{1, \ldots, M\}$, M standing for the number of cluster. For a given segmentation $\boldsymbol{\tau}$, we assume that

$$X_j | S_j = m \overset{i.i.d}{\sim} f_{\theta_m}(.), j \in [\tau_{k-1}+1, \tau_k].$$

$S_{\tau_{k-1}+1:\tau_k}$ being constant, we can introduce the random variable Z_k standing for the state of segment k with $k = 1, \ldots K$. Z_k is assumed to follow a multinomial distribution specified by its probability vector $(\pi_1, \ldots, \pi_M)$. The estimation of this model is based on an iterative procedure. K being given, the initialization step consists in running the DP algorithm to obtain the best segmentation in K segments. The two following steps are then repeated until convergence:

- For a given segmentation, the model parameters $(\theta_m, m = 1, \ldots M)$ and $(\pi_1, \ldots, \pi_M)$ are estimated using an EM algorithm.
- The parameters being fixed, the DP algorithm identifies the best segmentation in K segments.

As part of Rémi Patin's PhD thesis and in collaboration with Emilie Lebarbier and Simon Benhamou, we proposed a straightforward extension of this model to multivariate signals [**Pat+192**] and implemented it in the corresponding R package [**Pat+191**]. Although this approach assumed a very simplistic movement model (we chose a normal multivariate

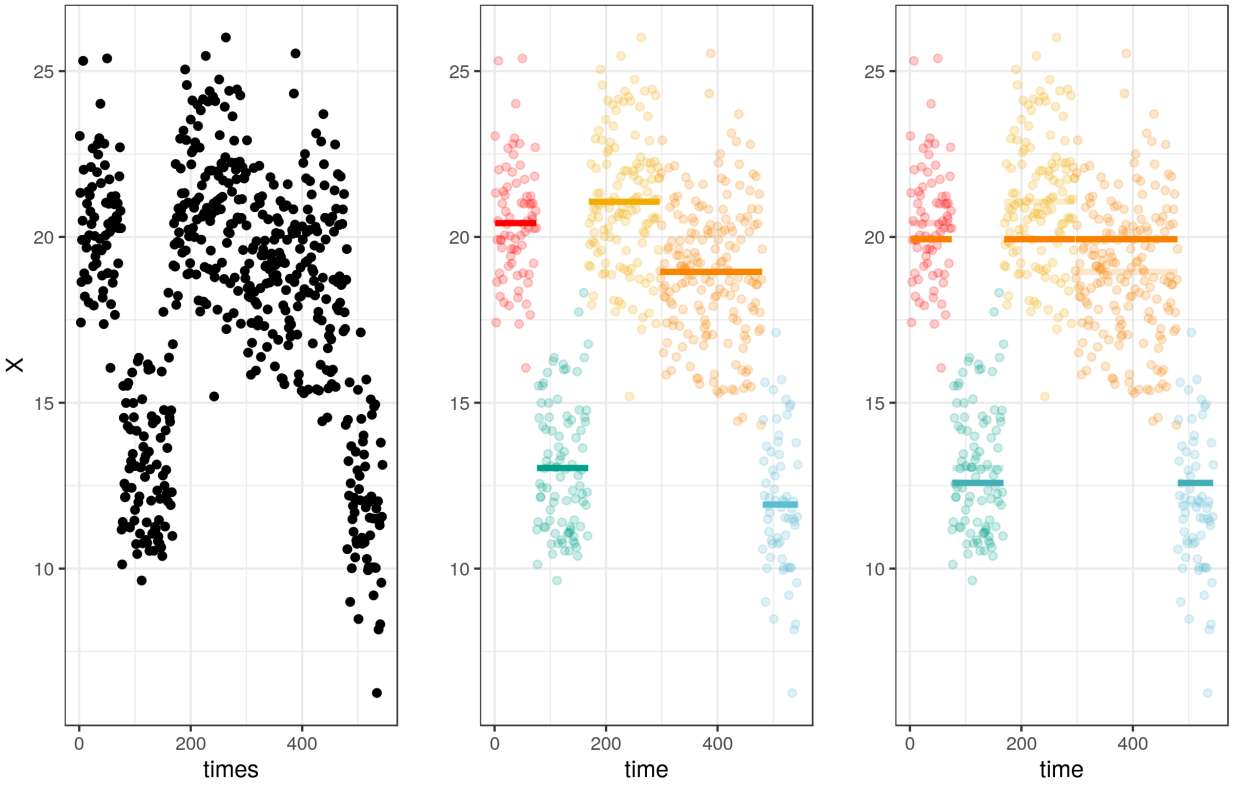


Figure 3.10: The left panel presents the a raw univariate signal to be segmented. The middle panel shows the result of the change point detection procedure on this signal. There are $K = 5$, three of them with war colors correspond to high expected values while the two other segment are characterized by smaller expected values. Finally the right panel, presents the segmentation-clustering model applied to this signal. There are two clusters, the first one consists of all high mean segments, while the second cluster is composed with the two low mean segments.

model), it has been proven to be very robust on simulations and real case applications. The main drawback relies on the modified BIC criterion we proposed to select the number of cluster which do not provide relevant biological results. This problem has also been mentioned by [Poh+17].

HMM methods and change point detection methods have first been developed for signal processing and therefore are part of classical toolkit for time series analysis. However, animal movement is not only a temporal object and we should also consider the spatial component.

I am currently developing with Julien Chiquet, Sophie Donnet and Adeline Samson an extension of the segmentation/clustering approach to a larger choice of movement models, which are treated as spatio-temporal object. When the likelihood is expressed as a sum over all segments, the optimization over the segmentation space can be handled using the DP algorithm. As such, most of classical movement models can be included in this approach. The limit can arise from the estimation step for a given segmentation, which should be quick enough to be treated within this DP approach to avoid prohibitive computation time. Brownian motion, Ornstein Uhlenbeck model, Continuous correlated random walk observed at some discrete times can be represented as linear models and, as so, the estimation step in straightforward. More interestingly, the Langevin model who explicitly includes the effects of spatial covariates on the movement, presented in section 3.3.5, thanks to the Euler approximation, can be approximated by a linear model. Coupling the segmentation-clustering approach to the Langevin model, we will be able to propose quite realistic movement models which account for switch in behaviors and to produce the utilization distribution map associated to each behavior but also the average utilization map by integrating the time spent in each behavior. This work as well as other extensions will be again discussed in the final conclusion of this document.

3.5 Conclusion

In this chapter, I have presented my recent work which aim at proposing efficient and realistic models and methods to study animal movement. At the end of the last section I drafted some on going work on this subject.

There are also some perspectives I would like to mention now. First it will be straightforward to include some random effects in the model presented in Equation Eq 3.3.8 to account for individual heterogeneity in the space utilization. Again thanks to the Euler approximation, the estimation of this random effect movement model will be reduced to the parameter estimation in a mixed effect model and could be included in the change point detection approach described in section 3.4.

The question of multiple trajectories is of great interest since the GPS device become cheaper every day, the number of individuals monitored simultaneously tends to increase everyday and this raises the question of joint movement analysis. This question is currently poorly addressed. The definition and the measure of joint movement are even not precisely defined, as we discovered while working on a review for measuring dyadic movement [**Joo+18**]. To my knowledge, joint movement analysis mostly rely on Individual Based model which attempts to reproduce emergent collective behavior from simple rules

between individuals. However Brillinger, Preisler, and Wisdom [BPW11] proposed a similar approach using SDE. For any individual i, $(i = 1, \ldots, N)$, they assumed that:

$$\mathrm{d}X_t^i = \nabla P(X_t^i)\,\mathrm{d}t - \nabla V(X_t^i, X_t^{-i})\,\mathrm{d}t + \gamma\,\mathrm{d}W_t, \quad \text{and } X_0^i = x_0^i$$

the function V being used to introduce interactions between individuals. This approach seems promising even if the interactions are not systematically instantaneous, especially in the case of leader-follower interactions. To account for switching behaviors, this model can be embedded in a hierarchical approach as long as efficient algorithms can be developed to estimate them.

Chapter bibliography

[AK+17] Patrik Andersson, Arturo Kohatsu-Higa, et al. "Unbiased simulation of stochastic differential equations using parametrix expansions". In: *Bernoulli* 23.3 (2017), pp. 2028–2057.

[And82] D John Anderson. "The Home Range: A New Nonparametric Estimation Technique: Ecological Archives E063-001". In: *Ecology* 63.1 (1982), pp. 103–112.

[Avg+16] Tal Avgar, Jonathan R Potts, Mark A Lewis, and Mark S Boyce. "Integrated step selection analysis: Bridging the gap between resource selection and animal movement". In: *Methods in Ecology and Evolution* 7.5 (2016), pp. 619–630.

[Bak+15] Laurie L Baker, Joanna E Mills Flemming, Ian D Jonsen, Damian C Lidgard, Sara J Iverson, and W Don Bowen. "A novel approach to quantifying the spatiotemporal behavior of instrumented grey seals used to sample the environment". In: *Movement Ecology* 3.1 (2015), p. 20.

[BB88] Pierre Bovet and Simon Benhamou. "Spatial analysis of animals' movements using a correlated random walk model". In: *Journal of Theoretical Biology* 131.4 (1988), pp. 419–433.

[**Beh+21**] Faustinato Behivoke, Marie-Pierre Etienne, Jérôme Guitton, Roddy Michel Randriatsara, Eulalie Ranaivoson, and Marc Léopold. "Estimating fishing effort in small-scale fisheries using GPS tracking data and random forests". In: *Ecological Indicators* 123 (2021), p. 107321.

[Ben11] Simon Benhamou. "Dynamic approach to space and habitat use based on biased random bridges". In: *PloS one* 6.1 (2011), e14592.

[Ben14] Simon Benhamou. "Of scales and stationarity in animal movements". In: *Ecology Letters* 17.3 (2014), pp. 261–272.

[Bes+06] Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O Roberts, and Paul Fearnhead. "Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.3 (2006), pp. 333–382.

[Bez+11] Nicolas Bez, Emily Walker, Daniel Gaertner, Jacques Rivoirard, and Philippe Gaspar. "Fishing activity of tuna purse seiners estimated from vessel monitoring system (VMS) data". In: *Canadian Journal of Fisheries and Aquatic Sciences* 68.11 (2011), pp. 1998–2010.

[**Bez+21**] Nicolas Bez, Marie-Pierre Etienne, Pierre Gloaguen, and Stéphanie Mahévas. "Evaluating Markov state space models' performances on annotated trajectories: simulation-estimation experiments and real cases." 2021.

[BPW11] DR Brillinger, HK Preisler, and MJ Wisdom. "Modelling particles moving in a potential field with pairwise interactions and an application". In: *Brazilian Journal of Probability and Statistics* (2011), pp. 421–436.

[Bri+02] David R Brillinger, Haiganoush K Preisler, Alan A Ager, John G Kie, and Brent S Stewart. "Employing stochastic differential equations to model wildlife motion". In: *Bulletin of the Brazilian Mathematical Society* 33.3 (2002), pp. 385–408.

[BSL+08] David R Brillinger, Brent S Stewart, Charles L Littnan, et al. "Three months journeying of a Hawaiian monk seal". In: *Probability and Statistics: Essays in honor of David A. Freedman*. Institute of Mathematical Statistics, 2008, pp. 246–264.

[Bur43] William Henry Burt. "Territoriality and home range concepts as applied to mammals". In: *Journal of Mammalogy* 24.3 (1943), pp. 346–352.

[Cag+10] Francesca Cagnacci, Luigi Boitani, Roger A Powell, and Mark S Boyce. *Animal ecology meets GPS-based radiotelemetry: A perfect storm of opportunities and challenges*. 2010.

[Cag+16] Francesca Cagnacci, Stefano Focardi, Anne Ghisla, Bram Van Moorter, Evelyn H Merrill, Eliezer Gurarie, Marco Heurich, Atle Mysterud, John Linnell, Manuela Panzacchi, et al. "How many routes lead to migration? Comparison of methods to assess and characterize migratory movements". In: *Journal of Animal Ecology* 85.1 (2016), pp. 54–68.

[Cal06] C. Calenge. "The package adehabitat for the R software: tool for the analysis of space and habitat use by animals". In: *Ecological Modelling* 197 (2006), p. 1035.

[CL63] William W Cochran and Rexford D Lord Jr. "A radio-tracking system for wild animals". In: *The Journal of Wildlife Management* (1963), pp. 9–24.

[CMR06] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

[DGA00] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering". In: *Statistics and Computing* 10.3 (2000), pp. 197–208.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[Dou+11] Randal Douc, Aurélien Garivier, Eric Moulines, Jimmy Olsson, et al. "Sequential Monte Carlo smoothing for general state space hidden Markov models". In: *The Annals of Applied Probability* 21.6 (2011), pp. 2109–2145.

[**EGM20**] Marie-Pierre Etienne, Pierre Gloaguen, and Théo Michelot. *Rhabit: R for habitat selection using movement data*. R package version 0.1.0. 2020. URL: https://github.com/papayoun/Rhabit.

[**Eti+21**] Marie-Pierre Etienne, Pierre Gloaguen, Sylvain Le Corff, and Jimmy Olsson. "Backward importance sampling for partially observed diffusion processes". In: *arXiv preprint arXiv:2002.05438* (2021).

[FC17] Christen H Fleming and Justin M Calabrese. "A new kernel density estimator for accurate home-range and species-range area estimation". In: *Methods in Ecology and Evolution* 8.5 (2017), pp. 571–579.

[Fea+17] Paul Fearnhead, Krzystof Latuszynski, Gareth O Roberts, and Giorgos Sermaidis. "Continuous-time importance sampling: Monte Carlo methods which avoid time-discretisation error". In: *arXiv preprint arXiv:1712.06201* (2017).

[Fle+15] Chris H Fleming, William F Fagan, Thomas Mueller, Kirk A Olson, Peter Leimgruber, and Justin M Calabrese. "Rigorous home range estimation with movement data: a new autocorrelated kernel density estimator". In: *Ecology* 96.5 (2015), pp. 1182–1188.

[FPR08] Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O Roberts. "Particle filters for partially observed diffusions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4 (2008), pp. 755–777.

[Fre+08] Carla Freitas, Christian Lydersen, Michael A Fedak, and Kit M Kovacs. "A simple new algorithm to filter marine mammal Argos locations". In: *Marine Mammal Science* 24.2 (2008), pp. 315–325.

[GAL09] Eliezer Gurarie, Russel D Andrews, and Kristin L Laidre. "A novel method for identifying behavioural changes in animal movement data". In: *Ecology Letters* 12.5 (2009), pp. 395–408.

[GDW04] Simon J Godsill, Arnaud Doucet, and Mike West. "Monte Carlo smoothing for nonlinear time series". In: *Journal of the American Statistical Association* 99.465 (2004), pp. 156–168.

[**GEL181**] Pierre Gloaguen, Marie-Pierre Etienne, and Sylvain Le Corff. "Online sequential Monte Carlo smoother for partially observed diffusion processes". In: *EURASIP Journal on Advances in Signal Processing* 2018.1 (2018), p. 9.

[**GEL182**] Pierre Gloaguen, Marie-Pierre Etienne, and Sylvain Le Corff. "Stochastic differential equation based on a multimodal potential to model movement data in ecology". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.3 (2018), pp. 599–619. DOI: 10.1111/rssc.12251.

[**Glo+15**] Pierre Gloaguen, Stéphanie Mahévas, Etienne Rivot, Matthieu Woillez, Jérôme Guitton, Youen Vermard, and Marie-Pierre Etienne. "An autoregressive model to describe fishing vessel movement and activity". In: *Environmetrics* 26.1 (2015), pp. 17–28.

[Glo15] Pierre Gloaguen. "Modélisation mécaniste et stochastique des trajectoires pour l'halieutique, Mechanistic and stochastique modelling of trajectories for fisheries science". PhD thesis. Université Européenne de Bretagne, Agrocampus-Ouest, 2015.

[GSS93] Neil J Gordon, David J Salmond, and Adrian FM Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE proceedings F (radar and signal processing)*. Vol. 140. 2. IET. 1993, pp. 107–113.

[Gur+17] Eliezer Gurarie, Christen H Fleming, William F Fagan, Kristin L Laidre, Jesús Hernández-Pliego, and Otso Ovaskainen. "Correlated velocity models as a fundamental unit of animal movement: Synthesis and applications". In: *Movement Ecology* 5.1 (2017), p. 13.

[HHA+15] Ephraim M Hanks, Mevin B Hooten, Mat W Alldredge, et al. "Continuous-time discrete-space models for animal movement". In: *The Annals of Applied Statistics* 9.1 (2015), pp. 145–165.

[Hoo+10] Mevin B Hooten, Devin S Johnson, Ephraim M Hanks, and John H Lowry. "Agent-based inference for animal movement and selection". In: *Journal of Agricultural, Biological and Environmental Statistics* 15.4 (2010), pp. 523–538.

[Hoo+17] Mevin B Hooten, Devin S Johnson, Brett T McClintock, and Juan M Morales. *Animal Movement: Statistical Models for Telemetry Data*. CRC Press, 2017.

[Hor+07] Jon S Horne, Edward O Garton, Stephen M Krone, and Jesse S Lewis. "Analyzing animal movements using Brownian bridges". In: *Ecology* 88.9 (2007), pp. 2354–2363.

[JFM05] Ian D Jonsen, Joanna Mills Flemming, and Ransom A Myers. "Robust state–space modeling of animal movement data". In: *Ecology* 86.11 (2005), pp. 2874–2880.

[JL18] Devin S. Johnson and Josh M. London. *crawl: an R package for fitting continuous-time correlated random walk models to animal movement data*. 2018. DOI: 10.5281/zenodo.596464. URL: https://doi.org/10.5281/zenodo.596464.

[Joh+08] Devin S Johnson, Joshua M London, Mary-Anne Lea, and John W Durban. "Continuous-time correlated random walk model for animal telemetry data". In: *Ecology* 89.5 (2008), pp. 1208–1215.

[Joo+11] Rocio Joo, Sophie Bertrand, Alexis Chaigneau, and Miguel Niquen. "Optimization of an artificial neural network for identifying fishing set positions from VMS data: an example from the Peruvian anchovy purse seine fishery". In: *Ecological Modelling* 222.4 (2011), pp. 1048–1059.

[**Joo+18**] Rocio Joo, Marie-Pierre Etienne, Nicolas Bez, and Stéphanie Mahévas. "Metrics for describing dyadic movement: a review". In: *Movement Ecology* 6.1 (2018), p. 26.

[Kay+15] Roland Kays, Margaret C Crofoot, Walter Jetz, and Martin Wikelski. "Terrestrial animal tracking as an eye on life and planet". In: *Science* 348.6240 (2015), aaa2478.

[Kes97] Mathieu Kessler. "Estimation of an ergodic diffusion from discrete observations". In: *Scandinavian Journal of Statistics* 24.2 (1997), pp. 211–229.

[Kie+10] John G Kie, Jason Matthiopoulos, John Fieberg, Roger A Powell, Francesca Cagnacci, Michael S Mitchell, Jean-Michel Gaillard, and Paul R Moorcroft. "The home-range concept: are traditional estimators still relevant with modern telemetry technology?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1550 (2010), pp. 2221–2231.

[KLS12] Mathieu Kessler, Alexander Lindner, and Michael Sorensen. *Statistical Methods for Stochastic Differential Equations*. Chapman and Hall/CRC, 2012.

[Lav05] Marc Lavielle. "Using penalized contrasts for the change-point problem". In: *Signal Processing* 85.8 (2005), pp. 1501–1510.

[LBI15] DC Lidgard, WD Bowen, and SJ Iverson. *Data from: A novel approach to quantifying the spatiotemporal behavior of instrumented grey seals used to sample the environment*. 2015. DOI: doi:10.5441/001/1.910p0c20. URL: http://dx.doi.org/10.5441/001/1.910p0c20.

[Leb18] Emilie Lebarbier. *Contributions à la segmentation de processus*. 2018.

[Lel+13] Subhash R Lele, Evelyn H Merrill, Jonah Keim, and Mark S Boyce. "Selection, use, choice and occupancy: clarifying concepts in resource selection studies". In: *Journal of Animal Ecology* 82.6 (2013), pp. 1183–1191.

[Lev92] Simon A Levin. "The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture". In: *Ecology* 73.6 (1992), pp. 1943–1967.

[LL12] Oscar D Lara and Miguel A Labrador. "A survey on human activity recognition using wearable sensors". In: *IEEE Communications Surveys & tutorials* 15.3 (2012), pp. 1192–1209.

[Lon+09] Ryan A Long, Jonathan D Muir, Janet L Rachlow, and John G Kie. "A comparison of two modeling approaches for evaluating wildlife-habitat relationships". In: *The Journal of Wildlife Management* 73.2 (2009), pp. 294–302.

[Man+02] BFL Manly, Lyman McDonald, Dana Thomas, Trent L McDonald, and Wallace P Erickson. *Resource selection by animals: statistical design and analysis for field studies, Second Edition*. Kluwer Academic Publishers, Dordrecht, 2002.

[Mar+15] Marza Ihsan Marzuki, René Garello, Ronan Fablet, Vincent Kerbaol, and Philippe Gaspar. "Fishing gear recognition from VMS data to identify illegal fishing activities in Indonesia". In: *OCEANS 2015-Genova*. IEEE. 2015, pp. 1–5.

[Mat03] Jason Matthiopoulos. "The use of space by animals as a function of accessibility and preference". In: *Ecological Modelling* 159.2-3 (2003), pp. 239–268.

[MBM19] Théo Michelot, Paul G Blackwell, and Jason Matthiopoulos. "Linking resource selection and step selection models for habitat preferences in animals". In: *Ecology* 100.1 (2019).

[McC+14] Brett T McClintock, Devin S Johnson, Mevin B Hooten, Jay M Ver Hoef, and Juan M Morales. "When to be discrete: the importance of time formulation in understanding animal movement". In: *Movement Ecology* 2.1 (2014), p. 21.

[**Mic+19**] Théo Michelot, Marie-Pierre Etienne, Paul Blackwell, and Pierre Gloaguen. "The Langevin diffusion as a continuous-time model of animal movement and habitat selection". In: *Methods in Ecology and Evolution* (2019).

[Mil+06] Joshua J Millspaugh, Ryan M Nielson, Lyman McDonald, John M Marzluff, Robert A Gitzen, Chadwick D Rittenhouse, Michael W Hubbard, and Steven L Sheriff. "Analysis of resource selection using utilization distributions". In: *The Journal of Wildlife Management* 70.2 (2006), pp. 384–395.

[MLP16] Théo Michelot, Roland Langrock, and Toby A Patterson. "moveHMM: an R package for the statistical modelling of animal movement data using hidden Markov models". In: *Methods in Ecology and Evolution* 7.11 (2016), pp. 1308–1315.

[MMT93] BFL Manly, Lyman McDonald, and Dana Thomas. *Resource selection by animals: statistical design and analysis for field studies*. Chapman & Hall, London, 1993.

[Mor+04] Juan Manuel Morales, Daniel T Haydon, Jacqui Frair, Kent E Holsinger, and John M Fryxell. "Extracting more out of relocation data: building movement models as mixtures of random walks". In: *Ecology* 85.9 (2004), pp. 2436–2445.

[Nat+08] Ran Nathan, Wayne M Getz, Eloy Revilla, Marcel Holyoak, Ronen Kadmon, David Saltz, and Peter E Smouse. "A movement ecology paradigm for unifying organismal movement research". In: *Proceedings of the National Academy of Sciences* 105.49 (2008), pp. 19052–19059.

[Nor+13] Joseph M Northrup, Mevin B Hooten, Charles R Anderson Jr, and George Wittemyer. "Practical guidance on characterizing availability in resource selection functions under a use–availability design". In: *Ecology* 94.7 (2013), pp. 1456–1463.

[OB71] Eugene Pleasants Odum and Gary W Barrett. *Fundamentals of Ecology*. Vol. 3. Saunders Philadelphia, 1971.

[OS+11] Jimmy Olsson, Jonas Ströjby, et al. "Particle-based likelihood inference in partially observed diffusion processes using generalised Poisson estimators". In: *Electronic Journal of Statistics* 5 (2011), pp. 1090–1122.

[OW+17] Jimmy Olsson, Johan Westerborn, et al. "Efficient particle-based online smoothing in general hidden Markov models: the PaRIS algorithm". In: *Bernoulli* 23.3 (2017), pp. 1951–1996.

[Oza92] Tohru Ozaki. "A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach". In: *Statistica Sinica* (1992), pp. 113–135.

[Pat+10] Toby A Patterson, Bernie J McConnell, Mike A Fedak, Mark V Bravington, and Mark A Hindell. "Using GPS data to evaluate the accuracy of state–space methods for correction of Argos satellite telemetry error". In: *Ecology* 91.1 (2010), pp. 273–285.

[Pat+17] Toby A Patterson, Alison Parton, Roland Langrock, Paul G Blackwell, Len Thomas, and Ruth King. "Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges". In: *AStA Advances in Statistical Analysis* 101.4 (2017), pp. 399–438.

[**Pat+191**] Remi Patin, Marie-Pierre Etienne, Emilie Lebarbier, and Simon Benhamou. *segclust2d: Bivariate Segmentation/Clustering Methods and Tools*. R package version 0.2.0. 2019. URL: https://CRAN.R-project.org/package=segclust2d.

[**Pat+192**] Rémi Patin, Marie-Pierre Etienne, Emilie Lebarbier, Simon Chamaillé-Jammes, and Simon Benhamou. "Identifying stationary phases in multivariate time series for highlighting behavioural modes and home range settlements". In: *Journal of Animal Ecology* (2019).

[Pic+05] Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. "A statistical approach for array CGH data analysis". In: *BMC Bioinformatics* 6.1 (2005), p. 27.

[PM12] Roger A Powell and Michael S Mitchell. "What is a home range?" In: *Journal of Mammalogy* 93.4 (2012), pp. 948–958.

[Poh+17] Jennifer Pohle, Roland Langrock, Floris M van Beest, and Niels Martin Schmidt. "Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement". In: *Journal of Agricultural, Biological and Environmental Statistics* 22.3 (2017), pp. 270–293.

[Pyk78] Graham H Pyke. "Are animals efficient harvesters?" In: *Animal Behaviour* 26 (1978), pp. 241–250.

[Rab89] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[Rah+18] A Rahman, DV Smith, B Little, AB Ingham, PL Greenwood, and GJ Bishop-Hurley. "Cattle behaviour classification from collar, halter, and ear tag sensors". In: *Information Processing in Agriculture* 5.1 (2018), pp. 124–133.

[RH09] Christian Rutz and Graeme C Hays. *New Frontiers in Biologging Science*. 2009.

[RR98] Gareth O Roberts and Jeffrey S Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268.

[RT+96] Gareth O Roberts, Richard L Tweedie, et al. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363.

[Rus+11] Tommaso Russo, Antonio Parisi, Marina Prorgi, Fabrizio Boccoli, Innocenzo Cignini, Maurizio Tordoni, and Stefano Cataudella. "When behaviour reveals activity: Assigning fishing effort to métiers based on VMS data using artificial neural networks". In: *Fisheries Research* 111.1-2 (2011), pp. 53–64.

[RY13] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Vol. 293. Springer Science & Business Media, 2013.

[Sch+20] Amandine Schmutz, Laurence Chèze, Julien Jacques, and Pauline Martin. "A Method to Estimate Horse Speed per Stride from One IMU with a Machine Learning Method". In: *Sensors* 20.2 (2020), p. 518.

[SL16] Ulrike E Schlägel and Mark A Lewis. "Robustness of movement models: can models bridge the gap between temporal scales of data sets and behavioural processes?" In: *Journal of Mathematical Biology* 73.6-7 (2016), pp. 1691–1726.

[TCB14] Henrik Thurfjell, Simone Ciuti, and Mark S Boyce. "Applications of step-selection functions in ecology and conservation". In: *Movement Ecology* 2.1 (2014), p. 4.

[Ver+10] Youen Vermard, Etienne Rivot, Stéphanie Mahévas, Paul Marchal, and Didier Gascuel. "Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models". In: *Ecological Modelling* 221.15 (2010), pp. 1757–1769.

[WB10] Emily Walker and Nicolas Bez. "A pioneer validation of a state-space model of vessel trajectories (VMS) with observers' data". In: *Ecological Modelling* 221.17 (2010), pp. 2008–2017.

[WHJ18] Kenady Wilson, Ephraim Hanks, and Devin Johnson. "Estimating animal utilization densities using continuous-time Markov chain models". In: *Methods in Ecology and Evolution* 9.5 (2018), pp. 1232–1240.

[Wil52] DH Wilkinson. "The random element in bird 'navigation'". In: *Journal of experimental Biology* 29.4 (1952), pp. 532–560.

[Wor89] Brian J Worton. "Kernel methods for estimating the utilization distribution in home-range studies". In: *Ecology* 70.1 (1989), pp. 164–168.

[Zha+14] Zejun Zhang, James K Sheppard, Ronald R Swaisgood, Guan Wang, Yonggang Nie, Wei Wei, Naxun Zhao, and Fuwen Wei. "Ecological scale and seasonal heterogeneity in the spatial behaviors of giant pandas". In: *Integrative Zoology* 9.1 (2014), pp. 46–60.

[ZS07] Nancy R Zhang and David O Siegmund. "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data". In: *Biometrics* 63.1 (2007), pp. 22–32.

Chapter 4

Hierarchical Bayesian models for abundance monitoring

Contents

The conservation and the sustainable use of the oceans, seas and marine resources is the object of the 14th sustainable development objective. The International Union for Conservation of Nature's (IUCN) Red List of Threatened Species evaluates the extinction risk of thousands of species and subspecies and propose a classification into 8 categories (Extinct, Extinct in the wild, Critically endangered, Endangered, Vulnerable, Near threatened, Least concern, Data deficient). Over 116 000 species have been assessed for The IUCN Red List up to now. The scientific assessments are conducted by different organizations (ICES expert groups, NGOs) and propose to define abundance indices in order to fit dynamic models used for projections in the coming years. Therefore the first step in assessment of wildlife populations consists in defining relative abundance indices.

Abundance is obviously the result of unobserved complex spatio-temporal processes. Monitored abundance data are very diverse: scientific surveys, commercial catch in the context of fisheries and citizen science data, the latter two resulting from a complex sampling design. As a consequence, the abundance data present complex dependence pattern which

are the result of the spatial structure of the population and their habitats, the dynamic of the monitored population, or of the sampling process. The hierarchical framework, introduced in section 1.2.2 appears well suited to account for these complex dependencies.

4.1 Hierarchical Bayesian modeling

The terminology of Hierarchical Bayesian Model (HBM) is introduced by Berger [Ber85] and become popular in environmental and ecological science from the 1990s. As stated by Clark [Cla05], Hierarchical structure brings flexibility to the modeler and efficient algorithms have now been developed (Monte Carlo with Markov Chains, Roberts and Rosenthal [RR04], Sequential Monte Carlo [DGA00] and more recently Hamiltonian Monte Carlo [BG15] and all their variants). Additionally, the success of those hierarchical models is also explained by the development of softwares which propose a general implementation of those algorithms so that the modeler do not have to worry on the statistical inference (historically WinBUGS is the first one [Spi+03], but many others have proposed alternative stochastic algorithms).

4.1.1 Hierarchical model

Hierarchical models are characterized by the use of hidden variables as presented in Figure 3.5 and, following Berliner [Ber96] they are formalized in terms of three different layers:

- A data level which represents the observation/sampling process and specifies the probability distribution of observations given the underlying process.
- A hidden (latent) process level which represents the ecological mechanisms.
- A parameter level.

The switching movement models presented in section 3.4 and the HMM are examples of Hierarchical model. Hierarchical model is often used in stock assessment to represent complex biological dynamical process. The data level represents how the different observations are linked to this underlying process (observations are often several scientific abundance indices, commercial catch declarations which may target different stage of the population, with different gears). The use of this intermediate level process is a powerful way to represent the complex dependence between the different observations. As a very simple example, in [**Weg+16**], we studied the evolution of the population of sympatric Subantarctic (*Arctocephalus tropicalis*) and Antarctic fur seals (*A. gazella*) at Marion Island. The colonies on the island are counted either through direct count, through a capture-mark-recapture (CMR) count, or count from the cliffs for inaccessible beaches. Some beaches are counted with the two or three methods. Using hierarchical modeling, we developed an integrated approach which accounts for all source of data simultaneously and also estimates the probability of under detection for the direct count and the top cliff counts compared to a CMR approach. The corresponding DAG is given in Figure 4.1. Since the different observation process are linked to unknown number of pups on the beach, this model account for the complex dependence between the different observed counts on a given beach.

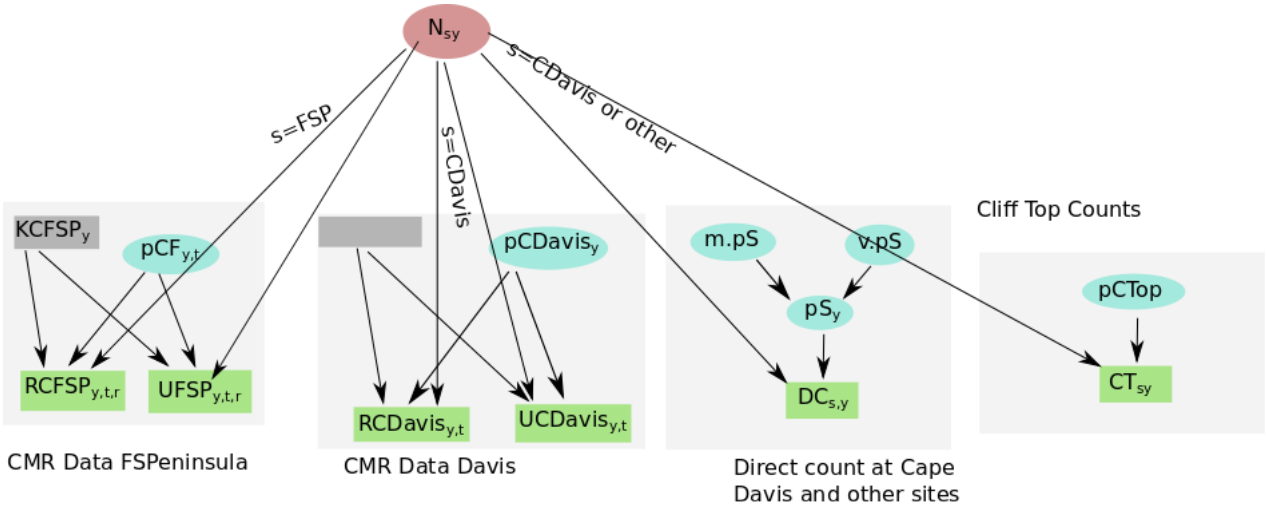


Figure 4.1: The directed Acyclic graph for the study of the pups seal populations on Marion Island. The green box represents the different dataset, the red oval is the unknown number of pup seals on beach s at year y and the blue ovals are the different parameters.

4.1.2 Bayesian inference

The flexibility offers by the hierarchical modeling approach is just one reason of the popularity of this framework among environmental modeler. In an article entitled *Why environmental scientists are becoming Bayesians*, Clark [Cla05] underlines the strength of the Bayesian approach especially in a decision-making context since uncertainty about parameters is propagated in predictions and inform the decision. The salmon stock assessment conducted by the Working Group on North Atlantic Salmon of the International Council for the Exploration of the Sea (ICES) uses a HBM at the Atlantic Ocean scale and integrates heterogeneous data (CMR data, survival data, commercial catches, ...) from sixteen countries [Ahl+19].

Even if the modeling flexibility offered by HBM partly explains the success of these approaches in ecology, it should not be forgotten that this success has been made possible by the development of high-performance softwares that offer an automatic implementation of recent estimation algorithms : WinBUGS, JAGS, INLA, Nimble, Stan [Spi+03; Plu03; MR09; Val+17; Car+17]. Thus, modelers are less concerned with parameter estimation issues, although a certain expertise in model parameterization remains essential.

Bayesian estimation is closely related to the notion of prior elicitation. The question of prior elicitation is of little importance in a data rich context, but most of ecological questions are addressed with few data because of the human cost, the monetary cost or the invasive side of data collection for most studies. Even in data rich stock management context, the increase in the biological realism of the models has produced complex models with parameters that are difficult to estimate with the available data. The question of prior elicitation then takes on its full importance. Prior knowledge can be defined thanks to expert knowledge [COM09] or meta analysis of the literature [**Sim+11**]. However the question of the prior choice remains a difficult and controversial issue. Whenever possible,

it is preferable to favor an objective Bayesian approach as defined Berger et al. [Ber+06]. Finally, to close this paragraph on the prior choice, I should mention that priors are also often chosen for their mathematical properties to improve the quality (in terms of mixing for example) of the numerical algorithm used for the estimation.

4.2 The question of the spatial representation of zero inflated data

As part of Sophie Ancelet PhD, I have been involved in a collaboration with Fisheries and Oceans Canada to analyze the abundance survey data from the southern Gulf of St. Lawrence (sGSL) figured by the red box in Figure 4.2. This survey consists on a stratified random design of scientific bottom-trawl surveys each September since 1971 . The stratification is based on depth and geographic area, with a varying number of sampled sites each year. The target fishing procedure is a 30mn straight-line tow at a given speed but it may vary depending on the weather conditions.

Abundance survey data are typical of zero inflated data as illustrated for the Urchin abundance in figure 4.2. This terminology refers to an excessive presence of zeros [TL14] and formally it corresponds to distribution whose mass at zero exceeds the expected mass at zero for any standard probability distribution function. Zero-inflated data are encountered in many disciplines, including agriculture, econometrics, epidemiology, industrial applications, public health. Zero inflated Count data has been the subject of many recent developments especially because of their use in metagenomics [Jon+19] or health care [CZZ19]. Until recently, continuous Zero inflated data has received less attention except in animal Ecology.

4.2.1 Zero inflated data

A classical approach consists in a so called two parts, hurdle or Delta models, which assume that zero and nonzero data arise, respectively, from separate processes [Ste96; MP04]. However the break between zero and nonzero values presents a particularly unnatural discontinuity in density data, where many zeros are actually stochastic clues of a strong gradient of decreasing biomass quantities as also discussed later by Thorson [Tho18]. Since the catches are the results of the towing process, a model which naturally accounts for the tow duration would be preferable.

In [**Anc+10**] and later in [**Lec+132**], we have explored the use of a compound Poisson with Gamma marks (a member of the Tweedie family) named CPG here after.

For a given sampling site, i, sampled during a duration D_i, the observed catch Y_i is defined by $Y_i \sim CPG(\mu, \alpha, \beta)$, i.e.

$$\begin{aligned} Y_i &= \sum_{l=0}^{N_s} Z_l, \quad Z_l \overset{i.i.d}{\sim} \Gamma(\alpha, \beta) \\ N_i &\sim \mathscr{P}(\mu D_i). \end{aligned} \tag{Eq 4.2.1}$$

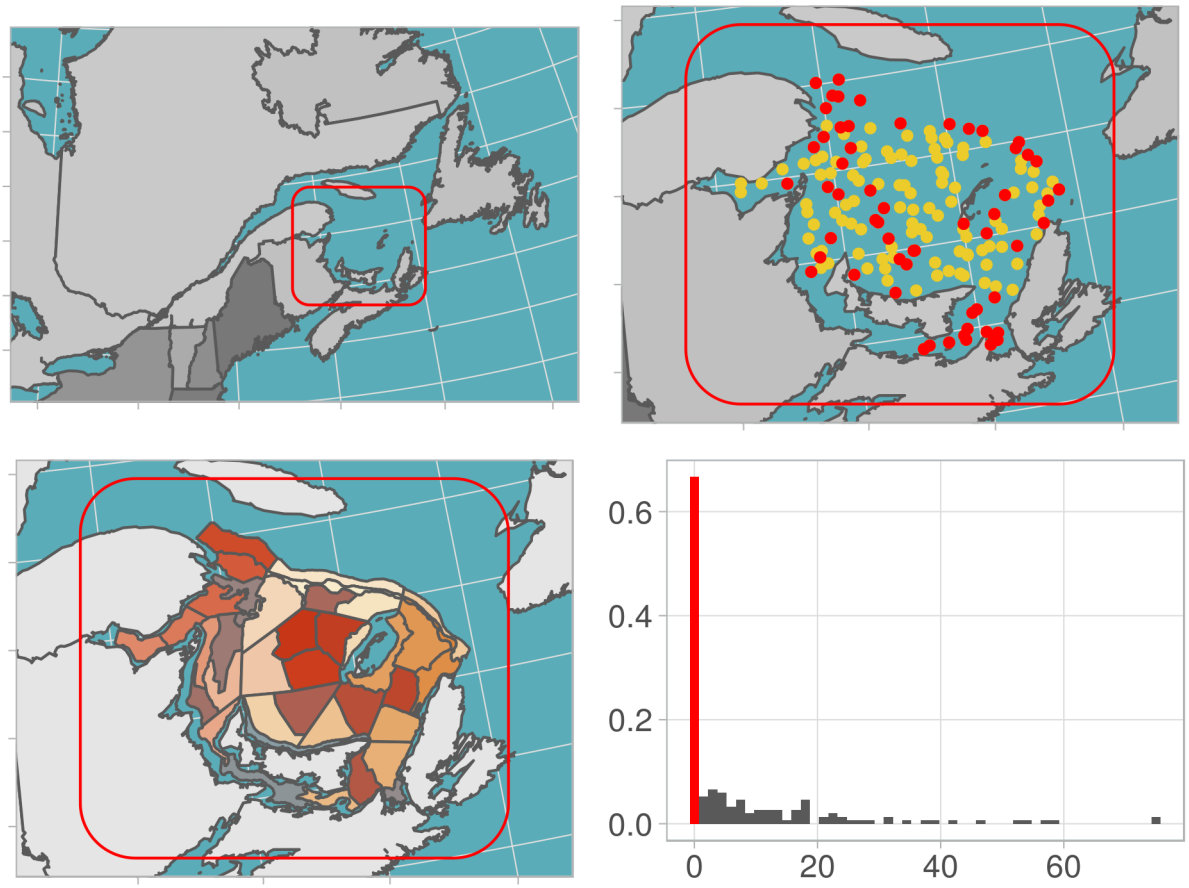


Figure 4.2: DFO conducts annual monitoring of invertebrate abundance in the southern Gulf of St. Lawrence figured by the red box. Focusing on the urchin biomass and year 1994, the red points indicate a zero catch while a yellow point indicates a positive observed biomass. The southern gulf is divided into 38 strata used for the stratified sampling design and quite homogeneous in terms of depth and habitat. The histogram represents the distribution of the urchin sampled weights.

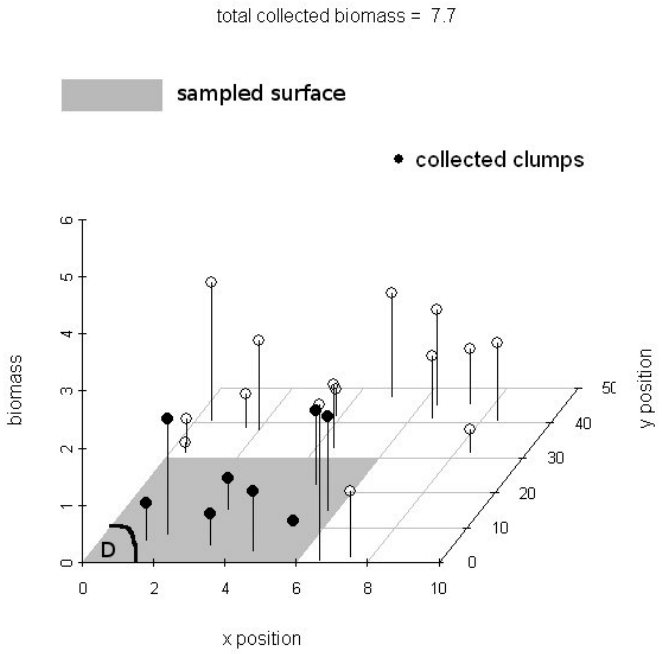


Figure 4.3: The compound Gamma (resp. Exponential) Poisson process defined as a compound Poisson process. The positions of clumps are drawn from a Poisson process and the biomass contained in a clump is also random with a Gamma distribution (resp Exponential). The towed are is figures by the grey zone. The observed biomass is then the sum of the biomass contained in every collected clumps.

From a hierarchical point of view, the model assumed that the clumps of biomass are distributed according to a Poisson process with intensity μ_s and the marks associated to the clumps (i.e. the biomass in every clumps) are Gamma distributed[1]. A schematic representation of the model is given in Figure 4.3.

The major advantage of this model is the additivity properties [Jør87] which provides a straightforward approach to incorporate the trawling duration by scaling the Poisson intensity parameter: the expected biomass is the product of the density μ_s at location s by the duration D_s[2]. In Delta models approach, the tow duration is either used to standardized the data or included as the covariates. In [**Lec+132**], we found that variable sampling volumes produce inference challenges for the Delta models but not for the Compound Poisson model which is consistent with the theoretical arguments we presented concerning the additivity

[1] In [**Anc+10**] , we focused on a simpler version where the marks are exponentially distributed.

[2] The correct value should be $\int_t^{t+D_s} \mu_{s(t)} dt$, where $s(t)$ is the path travelled by the vessel during the towing process. By approximating this quantity with $\mu_s D_s$, we assume that μ is almost constant at the tow level, that the vessel is moving at constant speed and we specify the intensity of the Poisson process up to a constant.

property.

The expected biomass recolted during a standard tow of duration D_{ref} is given by $\frac{\mu D_{ref}\alpha}{\beta}$ when the marks are distributed according to a Gamma distribution with shape α and rate β (the shape being equals to 1 for the exponential distribution).

4.2.2 Accounting for spatial dependence

The model proposed in Equation Eq 4.2.1 assumes that abundances is constant over the whole area. Thanks to hierarchical modeling framework, it is straightforward to propose more realistic abundance model.

We propose different improvements in [**Anc+10**] and [**Lec+131**] by adding a new process layer to model the spatial variation in the intensity of the Poisson process. As the scientific surveys of the Saint Lawrence Gulf is structured in homogeneous (in terms of habitat and depth) strata, we considered a regionalized version of the previous model, i.e a Cox process where the intensity of the Poisson process is constant for each strata μ_s, and $(\mu_s)s_{i=1,...S}$ are i.i.d. random gamma distributed variables. The Gamma distribution is chosen for its mathematical commodity to speed the estimation algorithm using partial conjugacy properties [3].

A major source of spatial organization arises from the spatial organization of the covariates and habitat preferences of different species. In [**Lec+131**], the intensity of the Poisson process is expressed as a linear function of covariates plus a spatial Gaussian Markov process using a logarithm link.

The modeling work developed for the analysis of scientific survey in the southern Saint Lawrence Gulf proved the flexibility of the compound Poisson process while used in a hierarchical modeling approach for analyzing spatially correlated zero inflated continuous data including covariates.

The estimation of those models have been conducted using different softwares (WinBUGS, JAGS and Nimble) but the computation time tends to become prohibitive with large dataset. The spatial process model could be expressed as a discrete grid convolution [Hig02; Cal04]. As part of his PhD project, Jean-Baptiste Lecomte explores the use of discrete grid convolution approach to reduce the problem dimension of this spatio-temporal model. Finding an accurate grid definition was found difficult at this time and the computation time was prohibitive except for some toys example.

In the last 10 years, the Stochastic Partial Differential Equations (SPDE) approach coupled with Integrated Nested Laplace Approximation [RMC09; LRL11; Kra+18] have been proven to be well suited for fitting complex models to many of the rich spatial data sets and the Tweedie distribution which generalizes the CPG model is now available in INLA.

[3] This choice permits a frequentist approach using the EM algorithm which is also available in a technical note in Etienne et al. [**Eti+09**]

4.3 On going and future work: Accounting for preferential sampling

The models presented above estimate abundance from spatially zero inflated data and implicitly assume that the sampling process and the abundance are independent conditionally on the accounted covariates. Although this assumption may be reasonable in the case of scientific survey, it is very questionable when it comes to data from commercial fisheries. Obviously the fishing vessels focus preferentially in areas with high potential, a situation refered as preferential sampling by Diggle, Menezes, and Su [DMS10], i.e. the sampled locations and the process of interest (here abundance density) are conditionally dependent given modeled covariates.

Commercial data are massive and cheap and represent a major source of information about abundance. For some fisheries there are even the major source of information (Atlantic tuna fisheries for example). Historically, commercial catches are corrected for different effects (zone, vessels, season, ...) through (generalized) linear models to build relative abundance indices. However, most methods ignore this preferential sampling issue.

Following Diggle, Menezes, and Su [DMS10], Pati, Reich, and Dunson [PRD11] propose a flexible model, spatially continuous, to account for preferential sampling when observations are assumed to follow a Gaussian distribution. More recently, Conn, Thorson, and Johnson [CTJ17] propose an extension to the previous approach where the observations are count data but consider discrete space (areal) models.

4.3.1 A first hierarchical model for preferential sampling

As part of Baptiste Alglave PhD thesis, I collaborate with Baptiste, Etienne Rivot (Agrocampus Ouest) , Youen Vermard (Ifremer) and Mathieu Woillez (Ifremer) to propose a preferential sampling model that accommodates zero inflated data. This model is intended to be used on the Sole fisheries to quantify how abundance varies in time.

The fish density is represented by a spatial field S defined by

$$S(x) = \exp\left\{\alpha_S + \sum_{i=1}^{d} \beta_d c_d(x) + \nu(x)\right\}, \quad \forall x \in D,$$

where

- D is the domain of interest,
- α_s a baseline constant,
- $c_d(x)$ is the value of the d^{th} environmental covariate at position x and β_d its effect,
- ν is a spatial Gaussian random field to account for spatial dependence.

Conditionally on S, the commercial sampling is then assumed to follow an inhomogeneous Poisson point process X with intensity λ, defined by:

$$\lambda(x) = \exp\{\alpha_C + \gamma \log S(x) + \eta(x)\}, \quad \forall x \in D,$$

where

- α_C is a baseline constant,
- and η is a spatial random field to account for the spatial dependence not represented by the fish density.

Conditionally on S and X, the commercial and scientific data are assumed to follow a log normal zero inflated distribution whose parameters are driven by S. The estimation of such hierarchical model can be challenging. However the direct maximization of an approximation of the log likelihood is possible by approximating all Gaussian random fields in the model by Gaussian Markov random fields as suggested in Lindgren, Rue, and Lindström [LRL11] and using Automatic differentiation [Nau11] as implemented in the tmb package [Kri+16]. As illustrated by a simulation study built to reflect the Sole fisheries data characteristics, the model behaves well. We showed that scientific data are not mandatory to provide good density estimation as soon as preferential sampling in commercial data is accounted for [**Alg+21**].

This model provides a solid basis for exploring the possibilities offered by the cheap and massive regulatory commercial data potentially available all year long (depending on the fishing restriction). Therefore we are now working to incorporate a temporal and a demographic component to this model.

4.3.2 Linking movement and catch data

Another question arises from the use of regulatory commercial catch data since the raw data consist in a weight of catch aggregated by species, spatial administrative unit, vessel and day. A map of spatial administrative units for the English Channel is produced in 4.4 as an example. Those area are quite large in regards to the potential spatial abundance variations in fish density.

Since most of the fishing vessels are now equipped with Vessel Monitoring System (see footnote 4, p. 53), an algorithm is used to identify the fishing locations (mostly a simple speed filter, but it could be any of the models presented in section 3.4). The catch are then proportionally allocated to the different fishing positions which results in geolocalized catches that are used to estimate abundance with the preferential sampling model presented above. However this two step procedure tends to produce overly smooth fish densities. Several research direction could be considered to improve the fish density estimation. First the previous model could be used to explore actual consequences of such proportional allocation on fish densities estimation. If the estimated density is proved to be sensitive to the choice of the allocation methods, the previous model should be updated to account for the aggregation of the data. Eventually, an integrated model linking fishing vessel movement and commercial catch data could be explored.

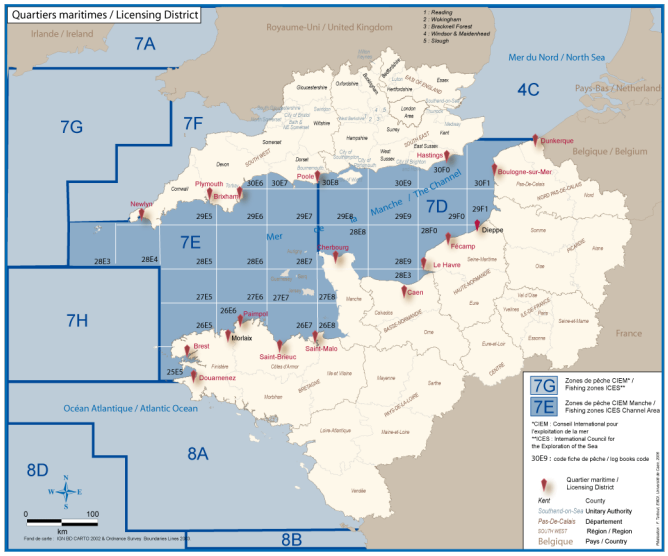


Figure 4.4: The ICEs fishing areas in the English Channel and the corresponding statistical units used for catch declarations (logbooks).

Chapter bibliography

[Ahl+19] Ida Ahlbeck-Bergendahl, Julien April, Hlynur Bardarson, Geir H Bolstad, Ian Bradbury, Mathieu Buoro, Gerald Chaput, Guillaume Dauphin, Dennis Ensing, Jaakko Erkinaro, et al. "Working group on North Atlantic salmon (WGNAS)". In: (2019).

[**Alg+21**] Baptiste Alglave, Youen Vermard, Marie-Pierre Etienne, Mathieu Woilez, and Etienne Rivot. "Integrated framework accounting for preferential sampling to infer fish spatial distribution". 2021.

[**Anc+10**] S. Ancelet, M.P. Etienne, H.P. Benoit, and E. Parent. "Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process". In: *Environmental and Ecological Statistics* 17.3 (2010), pp. 347–376. URL: http://www.springerlink.com/content/u40655u7361727q3/?p=7fe6ea1d569846b5928f845515d9a179&pi=1.

[Ber+06] James Berger et al. "The case for objective Bayesian Analysis". In: *Bayesian analysis* 1.3 (2006), pp. 385–402.

[Ber85] James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 1985.

[Ber96] L Mark Berliner. "Hierarchical Bayesian Time Series Models". In: *Maximum Entropy and Bayesian Methods*. Springer, 1996, pp. 15–22.

[BG15] Michael Betancourt and Mark Girolami. "Hamiltonian Monte Carlo for hierarchical models". In: *Current trends in Bayesian methodology with applications* 79.30 (2015), pp. 2–4.

[Cal04] Catherine A Calder. "Exploring latent structure in spatial temporal processes using process convolutions." PhD thesis. Duke University, 2004.

[Car+17] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1 (2017).

[Cla05] James S Clark. "Why environmental scientists are becoming Bayesians". In: *Ecology Letters* 8.1 (2005), pp. 2–14.

[COM09] Samantha Low Choy, Rebecca O'Leary, and Kerrie Mengersen. "Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models". In: *Ecology* 90.1 (2009), pp. 265–277.

[CTJ17] Paul B Conn, James T Thorson, and Devin S Johnson. "Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage". In: *Methods in Ecology and Evolution* 8.11 (2017), pp. 1535–1546.

[CZZ19] Tian Chen, Hui Zhang, and Bo Zhang. "A semiparametric marginalized zero-inflated model for analyzing healthcare utilization panel data with missingness". In: *Journal of Applied Statistics* 46.16 (2019), pp. 2862–2883.

[DGA00] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering". In: *Statistics and Computing* 10.3 (2000), pp. 197–208.

[DMS10] Peter J Diggle, Raquel Menezes, and Ting-li Su. "Geostatistical inference under preferential sampling". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59.2 (2010), pp. 191–232.

[**Eti+09**] Marie-Pierre Etienne, Eric Parent, Benoit Hugues, and Bernier Jacques. "Random effects compound Poisson model to represent data with extra zeros". In: *arXiv preprint arXiv:0907.4903* (2009).

[Hig02] Dave Higdon. "Space and space-time modeling using process convolutions". In: *Quantitative Methods for Current Environmental Issues*. Springer, 2002, pp. 37–56.

[Jon+19] Viktor Jonsson, Tobias Österlund, Olle Nerman, and Erik Kristiansson. "Modelling of zero-inflation improves inference of metagenomic gene count data". In: *Statistical Methods in Medical Research* 28.12 (2019), pp. 3712–3728.

[Jør87] Bent Jørgensen. "Exponential dispersion models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 49.2 (1987), pp. 127–145.

[Kra+18] Elias T Krainski, Virgilio Gómez-Rubio, Haakon Bakka, Amanda Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. *Advanced Spatial Modeling with Stochastic Partial differential Equations using R and INLA*. CRC Press, 2018.

[Kri+16] Kasper Kristensen, Anders Nielsen, Casper W. Berg, Hans Skaug, and Bradley M. Bell. "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software* 70.5 (2016), pp. 1–21. DOI: 10.18637/jss.v070.i05.

[**Lec+131**] J.B. Lecomte, H.P. Benoît, M.P. Etienne, L. Bel, and E. Parent. "Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical Bayesian model for zero-inflated biomass data". In: *Ecological Modelling* 265.0 (2013), pp. 74–84. ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2013.06.017.

[**Lec+132**] Jean-Baptiste Lecomte, Hugues P. Benoît, Sophie Ancelet, Marie-Pierre Etienne, Liliane Bel, and Eric Parent. "Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume". In: *Methods in Ecology and Evolution* 4.12 (2013), pp. 1159–1166. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12122.

[LRL11] Finn Lindgren, Håvard Rue, and Johan Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498.

[MP04] Mark N Maunder and André E Punt. "Standardizing catch and effort data: a review of recent approaches". In: *Fisheries Research* 70.2-3 (2004), pp. 141–159.

[MR09] Sara Martino and Håvard Rue. "Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program". In: *Department of Mathematical Sciences, NTNU, Norway* (2009).

[Nau11] Uwe Naumann. *The art of differentiating computer programs: an introduction to algorithmic differentiation*. SIAM, 2011.

[Plu03] Martyn Plummer. "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vol. 124. 125.10. Vienna, Austria. 2003, pp. 1–10.

[PRD11] Debdeep Pati, Brian J Reich, and David B Dunson. "Bayesian geostatistical modelling with informative sampling locations". In: *Biometrika* 98.1 (2011), pp. 35–48.

[RMC09] Håvard Rue, Sara Martino, and Nicolas Chopin. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series b (statistical methodology)* 71.2 (2009), pp. 319–392.

[RR04] Gareth O Roberts and Jeffrey S Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability surveys* 1 (2004), pp. 20–71.

[**Sim+11**] Maximilien Simon, Jean-Marc Fromentin, Sylvain Bonhommeau, Daniel Gaertner, and Marie-Pierre Etienne. *Investigating the performances of a bayesian biomass dynamic model with informative priors on Atlantic bluefin tuna*. Tech. rep. Collective Volume of Scientific Papers 66 (2), 811-828, 2011.

[Spi+03] David Spiegelhalter, Andrew Thomas, Nicky Best, and Dave Lunn. *WinBUGS user manual*. 2003.

[Ste96] Gunnar Stefánsson. "Analysis of groundfish survey abundance data: combining the GLM and delta approaches". In: *ICES Journal of Marine Science* 53.3 (1996), pp. 577–588.

[Tho18] James T Thorson. "Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative". In: *Canadian Journal of Fisheries and Aquatic Sciences* 75.9 (2018), pp. 1369–1382.

[TL14] Wanzhu Tu and Hai Liu. "Zero-inflated data". In: *Wiley StatsRef: statistics reference online* (2014), pp. 1–7.

[Val+17] Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. "Programming with models: writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics* 26.2 (2017), pp. 403–413.

[**Weg+16**] Mia Wege, Marie-Pierre Etienne, W Chris Oosthuizen, Ryan R Reisinger, Marthán N Bester, and PJ Nico De Bruyn. "Trend changes in sympatric Subantarctic and Antarctic fur seal pup populations at Marion Island, Southern Ocean". In: *Marine Mammal Science* 32.3 (2016), pp. 960–982.

Chapter 5

Conclusion

Contents

Every chapter of this document has ended with a short conclusion on the topic of interest. Therefore my purpose is not to repeat every points already mentioned before in this final conclusion. I use this chapter to highlight the way I conceive research in applied statistic for biology. As an applied statistician, I aim to develop new statistical methods useful for biology and ecology and this can't be achieved without understanding the questions raised by those disciplines. This is only possible because strong interactions with biologists and ecologists. Once the question is understood, its formalization is the first statistical work, which is then, of course, followed by a phase of theoretical or methodological development. One must then remain vigilant to deliver results that have a practical interest, and I detail what this means for me in the second part of this chapter. Finally, I will suggest some directions for my research to come.

5.1 Building close connections with practitioners

My PhD thesis was my first research experience at the interface between statistics and biology. My supervisor and I had identified a situation in the literature where the behavior of the local score was unknown. My work had consisted in exploring this behavior. However, when I tried to identify concrete situations to apply the results we had obtained, I found very few. Clearly, the chosen approach was not the right one.

It is obvious to me today that I must first immerse myself in biological issues to ensure that I develop relevant tools. However, after my PhD thesis, it took me quite a long time to

figure this out and identify a research theme in statistics that I thought would be useful. My sabbatical in 2009 in the Fisheries Center at the University of British Columbia, working with Dr Murdoch McAllister gave me the opportunity to take part in two stock assessments [**Yam+10**; **Yam+11**]. A few years later, I have been funded by the International Commission for the Conservation of Atlantic (ICCAT) to assess, in collaboration with Murdoch McAllister and Tom Carruthers, whether an age-structured model would improve the stock assessment of an iconic species, the Atlantic bluefin tuna (*Thunnus thynnus*). Contrary to what we had expected, we proved that given the available data and the associated uncertainties more complex models won't improve the assessment mostly because identifiability issues [**ECM13**][1].

Although this is not the core of my research activity, I try to maintain these collaborations in order to build or keep a link with colleagues interested in subjects other than statistics, specifically in my institution (AgroParisTech up to 2017, [**Gro+09**; **Sau+16**] and Agrocampus Ouest from then [**Gen+20**]). Most of the work developed during these collaborations might fall in the domain of statistical consulting but some pojects still require the development of ad-hoc models, as in [**Fri+16**] which involves a nice specific mixture models. Project students are also nice opportunities to maintain those collaborations.

5.2 Providing practical results that can be used by biologists and ecologists

Since the end of my PhD thesis, it is a constant concern that the research I am proposing should have practical outcomes. This concern can take different forms.

In my most theoretical works, I have always attempted to give explicitly computable results. When it involves a limit approximation, I have worked to bound the rate of convergence [**EV04**] as it is still currently the case with the rate of convergence of the pure jump Markov process to the Ornstein Uhlenbeck that my coauthors and I are striving to obtain. I have also evaluated the relevance of the so obtained bounds by conducting numerical simulations [**Eti02**] and this will be done as well to assess the performance of the proposed algorithm for the distribution of the longest excursion above a given threshold.

Continuous processes (continuous in time or in space) often prove to be useful in obtaining practically usable calculations. As an example, the exact formulas in finite horizon obtained by Daudin and Mercier [DM99] and recalled in formula Eq 2.1.4 for the local score, as well as the formula on recurrent alterations proposed by Robin and Stefanov [RS15] become too memory- and/or time-consuming as soon as long sequences are considered. On the contrary, limit approximations are very precise and are often easier to calculate[2]. In the context of movement ecology, utilization distribution and resource selection function based on movement data are often considered on discretized space [Avg+16] which requires heavy simulations. The Langevin model approach detailed in section 3.3.5 proves once again that

[1] Therefore Ben Bolker's talk at the International Statistical Ecology Conference held in 2014 [Bol14] discussing the concept of Statistical Machismo proposed by Brian McGill [McG12] had a particular resonance.

[2] Even if it still needs to be demonstrated for functionals involving the OU process.

continuous time formulation can sometimes circumvent time-consuming issues.

However continuous time models are not always an obvious solution. Accounting for observation error in movement model leads me to the study of partially observed diffusions. This practical question was at the origin of my interest on the estimation of POD. The works I have developed lead to contributions in statistics even so the proposed solutions are currently too computationally demanding to be of practical use. The question of the estimation of POD is a very active field of research in statistics and we can expect new improvement in the coming years which might produce more efficient algorithms.

Bringing statistical methods to practical use also means making them available in packages. I always gave access to my codes before I realized it wasn't nearly enough. The R software [R C18] provides a straightforward way to popularize statistical methods. Thanks to the impulse of two collaborators, I participated in the development of two packages recently [**gloaguen2020rhabit**; **Pat+191**], but I invest more in this aspect of the work in the years to come.

A key aspect of making a methodological development usable in practice involves publication in scope specific journals. This could be done in two ways. When the statistical method developed involved a bit of theory and is better suited to a statistical journal, it is important to collaborate to produce a practical application of this method. That was one of my goal with my PhD thesis work, but the studied assumptions seem to be too restrictive for practical use. I will continue to pursue theoretical work on genomic alterations. The other way consists in publishing in multidisciplinary journals which I now do as much as possible.

5.3 Perspective

This section proposes the perspective I would like to explore in the coming years. Some of them are short- or medium-term perspectives while others require a long term horizon.

5.3.1 Around the excursion of the Ornstein Uhlenbeck process

The work initiated around the convergence of a pure jump process to the Ornstein Uhlenbeck process and more specifically the convergence of the corresponding excursions is still in progress. As presented in the conclusion of the chapter 2, we have good directions to obtain the rate of convergence and then finish the corresponding paper. Nevertheless, many other questions still need to be explored around this subject. First the distribution of the lengths of excursions around m of an Ornstein Uhlenbeck process is not known except in the specific case where $m = 0$ [PY971]. As described in the last section of the same chapter, we propose an algorithm to compute the cumulative distribution function using a particle Monte Carlo algorithm. This algorithm relies on the simulation of the first hitting time to reach m, which is analytically given by either a series representation or an integral representation [APP05] but not so easy to sample from. Furthermore, as the the proposed algorithm will be used to evaluate the significance of an observed excursion, we expect to deal with large value of m, corresponding to the tail of the distribution. Consequently, we are lead to simulate rare

events with intractable CDF. There are many directions to explore and compare in order to provide practical results on the CDF of the longest excursion (such as simple discretization scheme or Sequential Monte Carlo for rare event simulation [Cér+12]).

5.3.2 Around the Langevin model for movement ecology

As mentioned in Chapter 3, the Langevin model used to link movement data to resource selection function is promising. Thank to a simple Euler approximation, this model can be simplified to a linear model.

This approximation opens many potential extensions. A first one would consist in accounting for individual heterogeneity by incorporating an individual random effects. From the estimation point of view, this would simply transform the Euler approximation into a mixed linear model instead of simple linear model. The resulting map could be examined at the population level for conservation management or at the individual level to study the space use heterogeneity among individuals. The Langevin model could also be embedded within a hierarchical approach to account for switch in resource selection function according to some internal unobserved states. A first simple version would assume, as proposed in section 3.4 that changes in behavior are synchronized with the sampled times. Unfortunately, this simple version would suffer from the classic drawbacks of discrete-time models which simultaneously model the movement process and the sampling process. To circumvent this assumption, we could assume that the state process is a continuous time pure jump process, but the simplicity of the Euler approximation would be lost. Conditionally on the switching times and using again the Euler approximation, the continuous time switching version of the Langevin model could be approximated by a mixture model, with some missing observations. Using uniformization approach [Ros+96] to sample the potential switches, we should be able to propose a Monte Carlo EM estimation algorithm.

Beyond these two well-identified aspects, I also plan to explore more open questions. The Langevin model proposed in [**Mic+19**] has been developed to account for quantitative environmental covariates, however suitable habitats is often described by discrete variables. First, I would like to explore a simple model, where the speed, characterized by the diffusion term, would change according to the visited habitat. This could be formalized as an extension to $\mathbb{R}^2$ of the results obtained by Lejay and Pigato [LP18]. Finally, I would be interested to explore the extension of the same model to account for dynamic covariates such as temperature, winds, etc ... The question has been raised many times while I was presenting the model and it is of major practical interest. As the movement is driven by the gradient of the environmental covariates at a given location, for a given time, the Langevin model could cope with dynamical covariates. However it opens many conceptual questions if only the definition of the stationary distribution in this context.

I believe that the perspectives described in this section can be used to define one or two interesting PhD thesis topics. According to the potential student interests and profiles, these subjects could be declined in more methodological (especially for the direct extensions of the Langevin model) or theoretical versions. A collaboration with Antoine Lejay is planned for the adaptation of these methods to movement models.

I conclude by emphasizing the aspects that are important in my approach to research. I particularly enjoy taking a problem from a field of application I am passionate about and formalizing it into a statistical problem. This work at the interface between disciplines requires an understanding of the major ideas of the field of application. It also requires being able to communicate simply on the mathematical aspects so as to discuss the relevance of the simplifying assumptions one is often led to make. I then take great pleasure in developing the necessary theoretical or methodological points. I am always surprised how simple applied questions might lead to sophisticated as well as useful statistical developments and questions. Finally, I place great emphasis on returning as much as possible to the field of application by providing tools or working collaboratively to communicate the obtained results .

Chapter bibliography

[APP05] L. Alili, P. Patie, and J. L. Pedersen. "Representations of the first hitting time density of an Ornstein-Uhlenbeck process". In: *Stochastic Models* 21.4 (2005), pp. 967–980. ISSN: 1532-6349. DOI: 10.1080/15326340500294702. URL: http://dx.doi.org/10.1080/15326340500294702.

[Avg+16] Tal Avgar, Jonathan R Potts, Mark A Lewis, and Mark S Boyce. "Integrated step selection analysis: Bridging the gap between resource selection and animal movement". In: *Methods in Ecology and Evolution* 7.5 (2016), pp. 619–630.

[Bol14] Ben Bolker. *Statistical machismo and common sense*. International Statistical Ecology Conference. July 2014. URL: https://www.slideshare.net/bbolker/montpellier-36611460.

[Cér+12] Frédéric Cérou, Pierre Del Moral, Teddy Furon, and Arnaud Guyader. "Sequential Monte Carlo for rare event estimation". In: *Statistics and Computing* 22.3 (2012), pp. 795–808.

[DM99] Jean-Jacques Daudin and Sabine Mercier. "Distribution exacte du score local d'une suite de variables indépendentes et identiquement distribuées". In: *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* 329.9 (1999), pp. 815–820.

[**ECM13**] Marie-Pierre Etienne, Tom Carruthers, and Murdoch K. McAllister. *Using Integrated Statistical Catch at age Model (iSCAM) for Atlantic Bluefin Tuna*. Tech. rep. ICCAT-GBYP 04/2013, 2013.

[**Eti02**] Marie-Pierre Etienne. "Le score local: un outil pour l'analyse de séquences biologiques". PhD thesis. Université Henri Poincaré-Nancy 1, 2002.

[**EV04**] Marie-Pierre Etienne and Pierre Vallois. "Approximation of the distribution of the supremum of a centred random walk. Application to the local score". In: *Methodology and Computing in Applied Probability* 6.3 (2004), pp. 255–275.

[**Fri+16**] Nicolas C Friggens, Christine Duvaux-Ponter, Marie-Pierre Etienne, Tristan Mary-Huard, and Philippe Schmidely. "Characterizing individual differences in animal responses to a nutritional challenge: towards improved robustness measures". In: *Journal of Dairy Science* To Appear (2016).

[**Gen+20**] Julien Genitoni, Danièle Vassaux, Alain Delaunay, Sylvie Citerne, Luis Portillo Lemus, Marie-Pierre Etienne, David Renault, Solenn Stoeckel, Dominique Barloy, and Stéphane Maury. "Hypomethylation of the aquatic invasive plant, Ludwigia grandiflora subsp. hexapetala mimics the adaptive transition into the terrestrial morphotype". In: *Physiologia Plantarum* 170.2 (2020), pp. 280–298.

[**Gro+09**] Sarah Groc, Jérome Orivel, Alain Dejean, Martin Jean-Michel, Marie-Pierre Etienne, Bruno Corbara, and Jacques H. C. Delabie. "Baseline study of the leaf-litter ant fauna in French Guianese forest". In: *Insect Conservation and Diversity* 2 (2009), pp. 183–193.

[LP18] Antoine Lejay and Paolo Pigato. "Maximum likelihood drift estimation for a threshold diffusion". In: *Scandinavian Journal of Statistics* (2018).

[McG12] Brian McGill. *Dynamic Ecology*. Sept. 2012. URL: https://dynamicecology.wordpress.com/2012/09/11/statistical-machismo/.

[**Mic+19**] Théo Michelot, Marie-Pierre Etienne, Paul Blackwell, and Pierre Gloaguen. "The Langevin diffusion as a continuous-time model of animal movement and habitat selection". In: *Methods in Ecology and Evolution* (2019).

[**Pat+191**] Remi Patin, Marie-Pierre Etienne, Emilie Lebarbier, and Simon Benhamou. *segclust2d: Bivariate Segmentation/Clustering Methods and Tools*. R package version 0.2.0. 2019. URL: https://CRAN.R-project.org/package=segclust2d.

[PY971] J. Pitman and M. Yor. "On the relative lengths of excursions derived from a stable subordinator". In: *Séminaire de Probabilités, XXXI*. Vol. 1655. Lecture Notes in Math. Berlin: Springer, 1997, pp. 287–305.

[R C18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[Ros+96] Sheldon M Ross, John J Kelly, Roger J Sullivan, William James Perry, Donald Mercer, Ruth M Davis, Thomas Dell Washburn, Earl V Sager, Joseph B Boyce, and Vincent L Bristow. *Stochastic Processes*. Vol. 2. Wiley New York, 1996.

[RS15] Stéphane Robin and Valeri T Stefanov. "Detection of significant genomic alterations via simultaneous minimal sojourns at a state by independent continuous-time markov chains". In: *Methodology and Computing in Applied Probability* 17.2 (2015), pp. 479–487.

[**Sau+16**] Ophélie Sauzet, Cécilia Cammas, Pierre Barbillon, Marie-Pierre Étienne, and David Montagne. "Illuviation intensity and land use change: Quantification via micromorphological analysis". In: *Geoderma* 266 (2016), pp. 46–57.

[**Yam+10**] Lynne K Yamanaka, Murdoch K. McAllister, Marie-Pierre Etienne, and Ron Haigh. *Stock Assessment for the inside population of yelloweye rockfish (Sebastes ruberrimus) in British Columbia, Canada for 2010*. Tech. rep. CSAP Working Paper 2010/P062010, 2010.

[**Yam+11**] Lynne K Yamanaka, Murdoch K. McAllister, Marie-Pierre Etienne, and Rob Flemming. *Stock Assessment and Recovery Potential Assessment for Quillback Rockfish (Sebastes maliger) on the Pacific Coast of Canada*. Tech. rep. CSAP Working Paper 2010/302011, 2011.

Appendix A

Some definitions and notations

A.1 Summary of the different symbols

A.1.1 Probabilistic symbols

$\mathbf{E}\{\}$	General notation for the expectation
$\mathbf{E}_\theta\{\}$	Expectation under parameter θ
$\mathbf{P}\{\}$	General notation for the probability
$\xrightarrow{(d)}$	Convergence in distribution
$\boldsymbol{W}$	A Brownian motion process
$\boldsymbol{W}^{\boldsymbol{\mu}}$	A Brownian motion with drift μ
$\boldsymbol{U}$	An Ornstein Uhlenbeck process

A.1.2 Generic mathematical symbol

$^\intercal$	Transposition symbol
$\mathscr{S}_2^+$	the set of 2×2 symmetric positive definite matrices
$vec(A)$	if A is a $m \times n$ matrix, $vec(A) = [a_{11}, \ldots, a_{m1}, a_{12} \ldots, a_{m2}, \ldots, a_{1n} \ldots, a_{mn}]^\intercal$
∇	Gradient operator
Δ	Laplace operator

A.2 Abreviations

cdf	Cumulative distribution function
iid	Independent and Identically Distributed
pdf	Probability distribution function
rv	Random variables
DAG	Directed Acyclic Graph
DP	Dynamic Programming
EM	Expectation Maximization
OU	Ornstein Uhlenbeck
POD	Partially Observed Diffusion

A.3 Operations

Definition 1 (Kronecker product)**.** Let A and B be 2×2 matrices, the Kronecker product is a 4×4 matrix defined by:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$$

Definition 2 (Kronecker sum)**.** Let A and B be 2×2 matrices, the Kronecker sum is defined by

$$A \oplus B = A \otimes I_2 + I_2 \otimes B$$

Appendix B

Scientific contribution

Book

[**Bel+15**] Liliane Bel, Jean-Jacques Daudin, Marie-Pierre Etienne, Emilie Lebarbier, Tristan Mary-Huard, Stéphane Robin, and Colette Vuillet. *Le modèle linéaire et ses extensions : modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences*. Paris: Ellipses Edition, 2015. ISBN: 978-2-340-00914-1.

Articles

[**Anc+10**] S. Ancelet, M.P. Etienne, H.P. Benoit, and E. Parent. "Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process". In: *Environmental and Ecological Statistics* 17.3 (2010), pp. 347–376. URL: http://www.springerlink.com/content/u40655u7361727q3/?p=7fe6ea1d569846b5928f845515d9a179&pi=1.

[**Beh+21**] Faustinato Behivoke, Marie-Pierre Etienne, Jérôme Guitton, Roddy Michel Randriatsara, Eulalie Ranaivoson, and Marc Léopold. "Estimating fishing ef-

fort in small-scale fisheries using GPS tracking data and random forests". In: *Ecological Indicators* 123 (2021), p. 107321.

[**BEM11**] Matias Braccini, Marie-Pierre Etienne, and Steve Martell. "Subjective judgement in data subsetting: implications for CPUE standardisation and stock assessment of non-target chondrichthyans". In: *Marine and Freshwater Research* 62 (2011).

[**Den+17**] Thomas Denis, Cécile Richard-Hansen, Olivier Brunaux, Marie-Pierre Etienne, Stéphane Guitet, and Bruno Hérault. "Biological traits, rather than environment, shape detection curves of large vertebrates in neotropical rainforests". In: *Ecological applications* 27.5 (2017), pp. 1564–1577.

[**DEV03**] Jean-Jacques Daudin, Marie-Pierre Etienne, and Pierre Vallois. "Asymptotic behavior of the local score of independent and identically distributed random sequences". In: *Stochastic Processes and their Applications* 107.1 (Sept. 2003).

[**EL13**] Marie-Pierre Etienne and Jean-Baptiste Lecomte. *La production Halieutique en Martinique - Avis d'expertise*. Tech. rep. DPMA - Ministère de l'Ecologie, 2013.

[**EV04**] Marie-Pierre Etienne and Pierre Vallois. "Approximation of the distribution of the supremum of a centred random walk. Application to the local score". In: *Methodology and Computing in Applied Probability* 6.3 (2004), pp. 255–275.

[**Fri+16**] Nicolas C Friggens, Christine Duvaux-Ponter, Marie-Pierre Etienne, Tristan Mary-Huard, and Philippe Schmidely. "Characterizing individual differences in animal responses to a nutritional challenge: towards improved robustness measures". In: *Journal of Dairy Science* To Appear (2016).

[**GEL181**] Pierre Gloaguen, Marie-Pierre Etienne, and Sylvain Le Corff. "Online sequential Monte Carlo smoother for partially observed diffusion processes". In: *EURASIP Journal on Advances in Signal Processing* 2018.1 (2018), p. 9.

[**GEL182**] Pierre Gloaguen, Marie-Pierre Etienne, and Sylvain Le Corff. "Stochastic differential equation based on a multimodal potential to model movement data in ecology". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.3 (2018), pp. 599–619. DOI: 10.1111/rssc.12251.

[**Gen+20**] Julien Genitoni, Danièle Vassaux, Alain Delaunay, Sylvie Citerne, Luis Portillo Lemus, Marie-Pierre Etienne, David Renault, Solenn Stoeckel, Dominique Barloy, and Stéphane Maury. "Hypomethylation of the aquatic invasive plant, Ludwigia grandiflora subsp. hexapetala mimics the adaptive transition into the terrestrial morphotype". In: *Physiologia Plantarum* 170.2 (2020), pp. 280–298.

[**Gim+14**] Olivier Gimenez, Stephen T Buckland, Byron JT Morgan, Nicolas Bez, Sophie Bertrand, Rémi Choquet, Stéphane Dray, Marie-Pierre Etienne, Rachel Fewster, Frédéric Gosselin, Bastien Mérigot, Pascal Monestiez, Juan M. Morales, Frédéric Mortier, François Munoz, Otso Ovaskainen, Sandrine Pavoine, Roger Pradel, Frank M. Schurr, Len Thomas, Wilfried Thuiller, Verena Trenkel, Perry de Valpine, and Eric Rexstad. "Statistical ecology comes of age". In: *Biology letters* 10.12 (2014), p. 20140698.

[**Glo+15**] Pierre Gloaguen, Stéphanie Mahévas, Etienne Rivot, Matthieu Woillez, Jérôme Guitton, Youen Vermard, and Marie-Pierre Etienne. "An autoregressive model to describe fishing vessel movement and activity". In: *Environmetrics* 26.1 (2015), pp. 17–28.

[**Gro+09**] Sarah Groc, Jérome Orivel, Alain Dejean, Martin Jean-Michel, Marie-Pierre Etienne, Bruno Corbara, and Jacques H. C. Delabie. "Baseline study of the leaf-litter ant fauna in French Guianese forest". In: *Insect Conservation and Diversity* 2 (2009), pp. 183–193.

[**Joo+18**] Rocio Joo, Marie-Pierre Etienne, Nicolas Bez, and Stéphanie Mahévas. "Metrics for describing dyadic movement: a review". In: *Movement Ecology* 6.1 (2018), p. 26.

[**Joo+21**] Rocio Joo, Nicolas Bez, Marie-Pierre Etienne, Pablo Marin, Nicolas Goascoz, and Stéphanie Mahévas. "Identifying 'partners at sea' on contrasting fisheries around the world". 2021.

[**Lec+131**] J.B. Lecomte, H.P. Benoît, M.P. Etienne, L. Bel, and E. Parent. "Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical Bayesian model for zero-inflated biomass data". In: *Ecological Modelling* 265.0 (2013), pp. 74–84. ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2013.06.017.

[**Lec+132**] Jean-Baptiste Lecomte, Hugues P. Benoît, Sophie Ancelet, Marie-Pierre Etienne, Liliane Bel, and Eric Parent. "Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume". In: *Methods in Ecology and Evolution* 4.12 (2013), pp. 1159–1166. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12122.

[**Mic+19**] Théo Michelot, Marie-Pierre Etienne, Paul Blackwell, and Pierre Gloaguen. "The Langevin diffusion as a continuous-time model of animal movement and habitat selection". In: *Methods in Ecology and Evolution* (2019).

[**Pat+192**] Rémi Patin, Marie-Pierre Etienne, Emilie Lebarbier, Simon Chamaillé-Jammes, and Simon Benhamou. "Identifying stationary phases in multivariate time series for highlighting behavioural modes and home range settlements". In: *Journal of Animal Ecology* (2019).

[**Per+18**] Marie Perrot-Dockès, Céline Lévy-Leduc, Julien Chiquet, Laure Sansonnet, Margaux Brégère, Marie-Pierre Étienne, Stéphane Robin, and Grégory Genta-Jouve. "A variable selection approach in the multivariate linear model: an application to LC-MS metabolomics data". In: *Statistical applications in genetics and molecular biology* 17.5 (2018).

[**Sau+16**] Ophélie Sauzet, Cécilia Cammas, Pierre Barbillon, Marie-Pierre Étienne, and David Montagne. "Illuviation intensity and land use change: Quantification via micromorphological analysis". In: *Geoderma* 266 (2016), pp. 46–57.

[**Sim+12**] Maximilien Simon, Jean-Marc Fromentin, Sylvain Bonhommeau, Daniel Gaertner, Jon Brodziak, and Marie-Pierre Etienne. "Effects of stochasticity in early life history on steepness and population growth rate estimates: an illustration on Atlantic bluefin tuna". In: *Plos ONE* 7.10 (2012).

[**Tys+20**] Niklas Tysklind, Marie-Pierre Etienne, Caroline Scotti-Saintagne, Alexandra Tinaut, Maxime Casalis, Valerie Troispoux, Saint-omer Cazal, Louise Brousseau, Bruno Ferry, and Ivan Scotti. "Microgeographic local adaptation and ecotype distributions: The role of selective processes on early life-history traits in sympatric, ecologically divergent Symphonia populations". In: *Ecology and Evolution* 10.19 (2020), pp. 10735–10753.

[**Weg+16**] Mia Wege, Marie-Pierre Etienne, W Chris Oosthuizen, Ryan R Reisinger, Marthán N Bester, and PJ Nico De Bruyn. "Trend changes in sympatric Subantarctic and Antarctic fur seal pup populations at Marion Island, Southern Ocean". In: *Marine Mammal Science* 32.3 (2016), pp. 960–982.

Package

[**EGM20**] Marie-Pierre Etienne, Pierre Gloaguen, and Théo Michelot. *Rhabit: R for habitat selection using movement data*. R package version 0.1.0. 2020. URL: https://github.com/papayoun/Rhabit.

[**Pat+191**] Remi Patin, Marie-Pierre Etienne, Emilie Lebarbier, and Simon Benhamou. *segclust2d: Bivariate Segmentation/Clustering Methods and Tools*. R package version 0.2.0. 2019. URL: https://CRAN.R-project.org/package=segclust2d.

Submitted papers

[**Alg+21**] Baptiste Alglave, Youen Vermard, Marie-Pierre Etienne, Mathieu Woilez, and Etienne Rivot. "Integrated framework accounting for preferential sampling to infer fish spatial distribution". 2021.

[**Bez+21**] Nicolas Bez, Marie-Pierre Etienne, Pierre Gloaguen, and Stéphanie Mahévas. "Evaluating Markov state space models' performances on annotated trajectories: simulation-estimation experiments and real cases." 2021.

Papers in preparation

[**Dec+211**] Laurent Decreusefond, Marie-Pierre Etienne, Gabriel Lang, and Stephane Robin. "An algorithm to compute the probability distribution of the excursionof Ornstein Uhlenbeck process". 2021.

[**Dec+212**] Laurent Decreusefond, Marie-Pierre Etienne, Gabriel Lang, Stéphane Robin, and Pierre Vallois. "Convergence of the sum of pure jump Markov unitary processes to an Ornstein Uhlenbeck process". 2021.

[**Eti+21**] Marie-Pierre Etienne, Pierre Gloaguen, Sylvain Le Corff, and Jimmy Olsson. "Backward importance sampling for partially observed diffusion processes". In: *arXiv preprint arXiv:2002.05438* (2021).

Technical and expert reports

[**Cad+18**] Noel Cadigan, Marie-Pierre Etienne, Mark Maunder, and Christian Reiss. *Summary Report CCAMLR Independent Stock Assessment Review for Toothfish*. Tech. rep. SC-CAMLR-XXXVII/02 Rev. 1, 2018.

[**ECM13**] Marie-Pierre Etienne, Tom Carruthers, and Murdoch K. McAllister. *Using Integrated Statistical Catch at age Model (iSCAM) for Atlantic Bluefin Tuna*. Tech. rep. ICCAT-GBYP 04/2013, 2013.

[**Eti+09**] Marie-Pierre Etienne, Eric Parent, Benoit Hugues, and Bernier Jacques. "Random effects compound Poisson model to represent data with extra zeros". In: *arXiv preprint arXiv:0907.4903* (2009).

[**Gio+08**] Jean-Marc Gion, Frédéric Mortier, Eric Mandrou, Hein Gherardi P.R, T Costecalde, Gilles Chaix, Marie-Pierre Etienne, P Sivadon, J. Grima-Pettenati, E. Villar, A. Saya, B. Pollet, C. Lapierre, and Vigneron P. "Un modèle de variabilité fonctionnelle chez les arbres forestiers : le gène CCR d'eucalyptus". In: *Actes du 7ème colloque national Ressources génétiques*. Oct. 2008.

[**Sim+11**] Maximilien Simon, Jean-Marc Fromentin, Sylvain Bonhommeau, Daniel Gaertner, and Marie-Pierre Etienne. *Investigating the performances of a bayesian biomass dynamic model with informative priors on Atlantic bluefin tuna*. Tech. rep. Collective Volume of Scientific Papers 66 (2), 811-828, 2011.

[**Yam+10**] Lynne K Yamanaka, Murdoch K. McAllister, Marie-Pierre Etienne, and Ron Haigh. *Stock Assessment for the inside population of yelloweye rockfish (Sebastes ruberrimus) in British Columbia, Canada for 2010*. Tech. rep. CSAP Working Paper 2010/P062010, 2010.

[**Yam+11**] Lynne K Yamanaka, Murdoch K. McAllister, Marie-Pierre Etienne, and Rob Flemming. *Stock Assessment and Recovery Potential Assessment for Quillback Rockfish (Sebastes maliger) on the Pacific Coast of Canada*. Tech. rep. CSAP Working Paper 2010/302011, 2011.

Résumé

Des nouvelles méthodes d'observations des organismes vivants font naître de nouvelles questions biologiques, de nouveaux types de données et un besoin de nouvelles méthodes d'analyse pour exploiter l'information qu'elles contiennent. Dans mon travail de recherche, je cherche à proposer et à étudier des modèles mathématiques pertinents pour répondre à des questions en génomique ou plus souvent en écologie. Ces modèles s'inscrivent principalement dans la famille des modèles Markoviens, potentiellement avec une structure cachée. Mes contributions en terme de recherche portent sur trois points essentiels. Une composante probabiliste qui se nourrit de problématique en génomique et qui m'a conduit à étudier la convergence de certains processus stochastique vers leur limite en temps continu. Un aspect statistique pour proposer des méthodes d'estimation, notamment dans le contexte de l'écologie du mouvement et plus spécifiquement pour des modèles définis par une équation différentielle stochastique partiellement observée. Une composante de modélisation pour proposer des modèles adaptés aux questions biologiques et aux types de données disponibles, notamment dans le cadre du suivi de la taille et de la répartition de population animale, exploitée ou non.

Abstract

New observation technics of living organisms give rise to new biological questions, new types of data and the need for the development of new statistical methods. My research aims to propose and study relevant mathematical models to answer questions in genomics or more often in ecology. These models belong most often to the Markovian models family, potentially built with a hidden structure. My contributions can be grouped into three main aspects.. A probabilistic component that feeds on genomic problems and that led me to study the convergence of certain stochastic processes to their limit in continuous time. A statistical aspect which consists in proposing estimation methods, particularly in the context of movement ecology and more specifically for models defined by a partially observed stochastic differential equation. A modeling activity to build relevant models for answering biological questions, given the available data in the context of population monitoring, whether exploited or not.