THÈSE DE DOCTORAT DE l'AGROCAMPUS OUEST

Effectuée à l'Institut Français de Recherche pour l'Exploitation de la Mer,
sous le label de l'Université Européenne de Bretagne,
pour obtenir le diplôme de :

DOCTEUR DE L'INSTITUT SUPERIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE

Spécialité : Écologie

École Doctorale : Vie Agro Santé

Présentée par

Pierre GLOAGUEN

Modélisation mécaniste et stochastique des trajectoires pour l'halieutique

devant le jury composé de :

Olivier Gimenez	Rapporteur
Adeline Leclercq Samson	Rapporteur
Didier Gascuel	Examinateur
Juan Manuel Morales	Examinateur
Samuel Soubeyrand	Examinateur
Stéphanie Mahévas	Directrice de thèse
Marie-Pierre Etienne	Co-directrice de thèse
Étienne Rivot	Co-directeur de thèse

Résumé

Impulsée par des questionnements sur les déterminismes du déplacement des individus, l'étude du mouvement en écologie s'est fortement développée ces dernières années. Cet engouement pour l'écologie du mouvement a été largement alimenté par l'émergence depuis 20 ans, de technologies GPS et par la constitution de nombreuses bases de données de trajectoires d'individus. Ces observations à des échelles spatiales et temporelles fines sont l'occasion de déceler les comportements des individus en lien avec l'environnement dans lequel ils évoluent. Pour reconstruire ces comportements et en comprendre les déterminismes sous-jacents, de nombreux modèles décrivant les trajectoires ont été développés et appliqués en écologie.

En parallèle à ce courant de recherche en écologie du mouvement, l'étude du mouvement et du comportement des navires de pêche a aussi connu un fort développement.. En effet, sous l'effet d'une réglementation européenne en vigueur depuis 2005, de nombreux navires de pêches européens sont équipés de système embarqués de surveillance (le Vessel Monitoring System, VMS), alimentant ainsi une importante banque de données de trajectoires. Ces données permettent d'aborder des questionnements variés, comme l'analyse de la dynamique spatio-temporelle de l'effort de pêche à fine échelle ou l'analyse du lien entre dynamique spatio-temporelle de l'activité de pêche et de la ressource. Ces données ont pour l'instant surtout été étudiées de manière descriptive, sans explicitation des mécanismes sous-jacents au déplacement. Ce travail de thèse vise à explorer la pertinence des modèles mécanistes, largement utilisés en écologie, pour répondre aux questions halieutiques.

Deux cadres de modélisation mécaniste et stochastique pour l'analyse des trajectoires sont développés et appliqués à l'analyse des trajectoires de navires de pêche. Tout au long de la thèse, un effort particulier est déployé pour le développement de méthodes d'inférence statistique basées sur la théorie du maximum de vraisemblance pour estimer les paramètres contrôlant le mouvement à partir d'observations discrètes des positions le long des trajectoires individuelles.

Le premier modèle vise à reconstruire le comportement de l'individu le long de sa trajectoire observée à pas de temps réguliers et repose sur des modèles de Markov cachés. Il est utilisé pour analyser les trajectoires de navires de pêche en Manche Est. Plus particulièrement, l'approche permet de déterminer les séquences d'activités de pêche différentes au cours d'une marée, et ainsi d'atteindre une description spatiale et temporelle de l'effort de pêche à fine échelle. La pertinence de ce modèle par rapport aux approches descriptives et mécanistes existantes est discutée. Cette approche a permis de mettre en évidence l'importance de prendre en compte les courants de marées dans l'étude des trajectoires de navires en Manche Est.

Le second cadre de modélisation propose une modélisation continue en temps et en espace et propose d'introduire un mécanisme explicite de dépendance de la trajectoire par rapport à l'environnement. Il repose sur les modèles d'équations différentielles stochastiques dans lequel

la part de déterminisme dans le mouvement (dérive) retranscrit l'idée de l'existence d'un champ spatial sous-jacent à la trajectoire qui représenterait des zones fortement attractives pour la pêche. Un algorithme d'estimation de ces zones à partir d'observations de positions (type VMS) est développé et appliqué aux navires de la Manche Est. Cet algorithme se base sur les avancées récentes pour la simulation exacte des processus de diffusion. Enfin, le modèle est utilisé pour tester une hypothèse d'importance en halieutique : les zones exploitées par les pêcheurs recoupent-t-elles les zones de forte abondance estimées par campagne scientifique ?

Au-delà des applications particulières sur lesquelles s'appuient ce travail de thèse, les modèles développés ont permis de construire des bases fondamentales prometteuses pour de nouveaux modèles en halieutique et en écologie du mouvement en général. De plus, les algorithmes d'inférence associés et explicités dans le manuscrit sont innovants et robustes à de nombreux ajouts pour un modèle plus réaliste.

Abstract

Driven by ecological questions about determinism of individuals, movement ecology has known large developments in recent years. This enthusiasm was largely possible thanks to the emergence of GPS technologies for the last 20 years, and, therefore, the establishment of numerous databases of individual trajectories. These observations at fine spatial and temporal scales of individuals give the opportunity to identify specific behaviors of individuals and their perception of the environment in which they operate. To process this data, and identify the mechanisms of interest to ecologists, many recent movement models describing the movement mecanisms have been developed.

In fisheries Science, with the development of GPS systems, the study of movement and behavior of fishing vessels has also experienced strong growth over the past ten years. Indeed many European fishing vessels are now mandatory equipped with onboard monitoring system (Vessel Monitoring System, VMS) creating a large monitored database. Primarly used for monitoring, these data are now used to address a variety of questions, such as analysis of the fishing effort at fine spatial and temporal scale of the fishing effort, or the analysis of the relationship between spatial-temporal dynamics of fishing activity and resource allocation.

These data are mostly studied descriptively, without explicitation of the mechanisms underlying the movement. This thesis aims to explore the relevance of mechanistic models, widely used in ecology to address fisheries issues.

Two stochastic and mechanistic modeling frameworks for analysis of trajectory data developed and applied to samples of the VMS database. Throughout the thesis, a special effort is being made to the development of statistical inference method to estimate the parameters controlling the movement. These inference is performed using discrete observations of positions along individual paths.

The first model is binding behavior and trajectory and is based on hidden Markov models. Assuming that a transformation of the speed process of an individual follows an autoregressive process, the model is used to estimate fishing activity from trajectory data. The approach therefore achieves a fine description of the fishing effort. The relevance of this model compared to existing descriptive and mechanistic approaches is discussed. Analyzing the performance of the model lead us to investigate the relevance of using speed relative to the water instead of speed relative to the water mass when studying fishing vessel trajectories. This study highlighted the importance of the surface currents in the study of fishing vessel trajectories in Eastern Channel.

The second approach develops a space and time continuous model that provides explicit mechanisms of drivers for trajectories. Individuals are supposed to follow the gradient of a cognitive map, representing their perception of the surrounding environment. The model is based on

stochastic differential equations in which the deterministic component (the drift) of the motion reflects subjective attractive areas for fishermen. An estimation algorithm of these areas from positions of observations is developed and applied to french fishing vessels in the English Channel. The model, and the resulting estimates are used to test an important hypothesis in fisheries science. Do areas used by fishermen overlap with areas of high abundance estimated by scientific surveys ?

The models developed here offer a promising basis for new models in fisheries and movement ecology. Moreover, associated inference algorithms detailled in the manuscript are robust to many extensions leading to a more realistic model.

Remerciements à mes financeurs

Ce travail de thèse a été effectué grâce à différents financements qui m'ont permis de travailler dans de très bonnes conditions durant 3 ans. J'ai ainsi pu bénéficier de l'apport de différentes sources :

- Une bourse IFREMER qui a financé la moitié ma thèse ;
- Une bourse de la région Pays de la Loire qui a financé l'autre moitié de ma thèse ;
- Le projet UE-FP7 VECTORS qui a financé de nombreux trajets pour des réunions et collaborations dans le cadre de ce groupe :
  FP7-VECTORS [2011-2014] "Vectors of Change in Oceans and Seas Marine Life" (7th EU Framework Program, Work program topic : OCEAN.2010-2 Vectors of changes in marine life, impact on economic sectors). Coordinateur : Melanie Austen (Univ. Portsmouth, UK) ;
- Le Groupement de Recherche Écologie Statistique, qui m'a octroyé une bourse finançant une très fructueuse collaboration avec l'AgroParisTech ;
- Le réseau Stats et Trajectoires (INRA) qui a financé de nombreuses et enrichissantes réunions sur le sujet de l'étude des trajectoires ;
- Le consortium européen EUR-OCEANS qui, au travers du groupe EtatJ'erre, a financé de nombreuses réunions autour de l'étude des trajectoires ;
- L'Université Européenne de Bretagne, qui m'a octroyé une bourse de mobilité m'a ainsi permis de séjourner durant 3 mois au National Marine Mammals Laboratory à Seattle ;
- Le SIH et le projet RECOPESCA, pilotés par Patrick Berhou et Emilie Leblond, qui, s'ils ne m'ont pas soutenu financièrement, m'ont permis d'accéder à un élément fondamental de cette thèse, les données VMS et RECOPESCA.

Je tiens ainsi à remercier vivement tous ces organismes, sans qui ce travail de thèse aurait été impossible.

Remerciements

Quatre années se sont écoulées depuis mon arrivée à IFREMER. Ces années ont été la suite de sept années d'études m'ayant mené sur diverses routes. Et maintenant arrive la difficile tâche de remercier tout les gens qui m'ont permis d'arriver au terme de cette thèse.
Pour moi cette thèse a été formidable sur le plan humain. J'ai eu la chance d'être encadré par trois personnes fantastiques tant humainement que scientifiquement. Être entouré de personnes aussi intelligentes, humbles et drôles a certainement été ma plus grande source d'inspiration durant ces quatre années.

Je tiens tout d'abord à remercier Stéphanie qui m'a supporté tout au long de ces quatre années à Nantes. En plus de subir mon entêtement et mes élucubrations, tu m'as toujours soutenu dans le travail qu'on a pu faire, en m'insufflant toujours une énergie positive qui te caractérise. Tu m'as toujours soutenu dans mes souhaits de différents projets et collaborations, et tu m'as encouragé quand je cédais à la lassitude. Je me souviens que tu m'as motivé quand je pensais inutile pour moi d'aller à Rochebrune, et comme la suite t'a donné raison ! Je te remercie pour les discussions scientifiques dans ton bureau, pour ta curiosité communicative, pour ta patience face à ma "tête de mule" [1], et pour ton éternel sourire ! Merci Steph !

Lorsque Steph n'en pouvait plus de moi, elle m'envoyait à Paris, chez Marie, ma tata lorraine ! Merci Marie pour ton accueil quand je venais à l'Agro. Merci pour ta patience dans ton enseignement, et pour tous les moments passés à discuter de maths, de stats, et d'autres choses de la vie. Merci pour tes encouragements et ta volonté qui m'ont poussé à m'attaquer aux EDS. Merci pour ta compréhension de ma personnalité et la confiance que tu m'as accordée pendant ma thèse et pour la suite. Merci pour ton sourire, et enfin merci pour l'accueil chez vous, avec Cyril et Maïa, et les soirées à jouer [2] avec les gens du labo ! Merci Marie !

Enfin, merci Étienne, le cachet "virilité" de mon encadrement. Merci pour tes discussions toujours justes et pertinentes, pour ton humilité à toute épreuve, pour tes encouragements incessants. Merci pour ton soutien continu dans les orientations prises durant la thèse, et la confiance que tu m'as donnée pour des interventions à Rennes. Un merci gigantesque pour l'aide à la rédaction du manuscrit, où tes commentaires écrits et téléphoniques m'ont constamment

1. Citation authentique
2. Et boire un peu aussi

relancé. Enfin un merci pour tes félicitations aux moments opportuns [3] et ta collaboration dans la "mission de secours" de Rochebrune ! Merci Tinou !

Parallèlement à l'encadrement officiel, je remercie ceux qui m'ont également encadré officieusement. Merci à toi Mathieu pour l'aide que tu m'as apportée durant mon stage et au début de ma thèse, bosser et partager le quotidien d'un post doc comme toi à été un sacré plaisir. Tu m'as appris à rester un véritable roc en toutes circonstances [4]. Merci à toi Youen [5] pour ton aide en halieutisme, ton aide concernant les données, ta bonne humeur et ton torse velu ! Merci à toi Sylvain, cette rencontre au sommet de Rochebrune a été le tournant de ma thèse. Ton intelligence et ta patience m'ont incroyablement aidé, ça a été pour moi une grande chance de faire la connaissance de quelqu'un d'aussi humble et fort [6] dans son domaine.
I'd like to thank you, Devin, for having welcomed me in Seattle. I spent three awesome months there, and I learnt a lot in few months in your lab. I hope our roads will cross again !

Je voudrais remercier toutes les personnes avec qui j'ai travaillé pendant près de quatre ans au laboratoire EMH. L'ambiance au labo a toujours été géniale, et la proximité de spécialistes en halieutique et en biologie m'a permis d'apprendre énormément sur des domaines inconnus pour moi. Un énorme merci [7] à Olivier, Informaticien parmi les informaticiens, superbe praticien du Linux, avant gardiste du langage R, humoriste de renom [8], et éternel animateur du labo. Merci pour ton aide informatique précieuse [9], notamment sur le très intuitif Caparmor. J'eusse aimé que tu sois en Islande avec moi pour y rencontrer mon fils [10].
Je voudrais remercier spécialement les thésards qui m'ont accueilli quand j'étais stagiaire, et avec qui j'ai passé des sacrés moments. Adri [11], pour toutes les soirées n'ayant eu qu'un seul objectif [12], et pour toutes les discussions scientifiques en halieutisme. Babwa [13], pour ta joie de vivre, ta superbe répartie, et ton ouverture d'esprit. Loïc [14], pour toutes les chouettes soirées à Nantes, les LDC [15] au labo, les roustes au Badminton et ta fidélité à notre lieu favori [16]. Laurence [17], pour les chouettes discussions, les tours à la voile, ta patience, et les bouteilles de vins. La Science a pris une saveur formidable grâce à vous !
Merci également spécialement à deux stagiaires, promo 2013, j'ai nommé Junior "Babyface"

3. Notamment ce coup de téléphone, un samedi brumeux, à Reykjavik
4. Absolument toutes
5. Monsieur Manche Est (MME)
6. LeCorff fort
7. Mersou !
8. Pas mal de blagues tirées par les cheveux, quand même
9. Tu aides la Science, et c'est ta Joie
10. Son nom est Pierreson
11. Tu bois
12. Rarement atteint
13. Ça s'écrit Bwbww ?
14. Artiste de cirque
15. IFREMER donne goût aux acronymes
16. Le Marlowe, évidemment
17. Rhabille toi...

Prod'homme [18] et Got Got [19], qui ont au moins appris la pétanque [20] à IFREMER, et c'est déjà ça.
Ma fin de thèse n'aurait pas été la même [21] sans tout le groupe arrivé en 2015. Merci aux "Dirty thésards", Riton [22], Xanxander [23] et Helmut [24], et bonne chance pour votre fin de thèse ! Et merci aux Inglorious Stagiaires, Benji la Malice [25], Tête de Fraise [26], Robaaaiiiinnnn [27], Tutur [28], Pumba [29] et Florianne [30] ! Vous avez tous très bien appris la coinche en étant chez nous, et c'est déjà énorme ! Bonne chance pour la suite !
Parmi les permanents, un immenze Merzi à Zigfried [31], zinzèrement, z'était zuper zympa de bozzer avec toi. Merci à Vincent pour toutes les discussions au café, au repas, ou encore sur le Gwen Dreizh. Merci à Mathieu pour les très instructives discussions sur l'acoustique. Merci à Anne Sophie pour l'éternelle bonne humeur et son goût du LaTeX. Merci à Jacques pour son splendide pot de départ en retraite [32], ses histoires de mer, et sa manière d'expliquer la science à la télé. Merci à Pascal L. pour avoir accepter de m'amener sur Febbe et m'avoir fait découvrir quelque chose d'incroyable : les poissons. Merci à Pascal L. pour m'avoir montré que mon style vestimentaire pouvait perdurer avec l'âge. Merci à Polo pour m'avoir aiguillé vers la boule Nantaise. Un merci également à Anik, Manuella, Verena, Marie-Joëlle, Aurélie, Emmanuelle et Fabien. Un grand merci à Anne et Isabelle qui n'ont jamais perdu courage à m'expliquer ce qu'était une formalité administrative. Enfin merci au Big Boss d'EMH (BB EMH), Pierre [33], pour les innombrables discussions de science qui m'ont énormément servi, pour ton temps précieux que tu m'as accordé pour assouvir ma curiosité, pour ta curiosité sur mes travaux, pour les balades à la voile dans la baie de Bourgneuf, pour ton enthousiasme permanent, et pour les souvenirs à tourner la grande roue de notre fortune à Reykjavik.

Je remercie également provincialement tous les gens de l'AgroParisTech qui m'ont super bien accueilli quand j'ai débarqué là bas. Je pense notamment à Fred, Aurélien, Pierre B., Anna, Loïc, Marie C., Maud, Julien, Vincent et Tristan. J'ai découvert grâce à vous un labo de gens très intelligents, cultivés, enjoués et drôles [34]. J'ai hâte d'arriver pour de nouveaux quizz de l'agro [35] ! Un grand merci également à Eric, qui a accepté d'être mon tuteur de thèse, et dont

18. Tu grandis à une vitesse...
19. Reste humble...
20. Sur le plus grand terrain de pétanque d'Europe ©
21. En effet, ça aurait été bouclé 3 mois plus tôt
22. Tu te présentes, tu t'appelles Henri
23. C'est la fête du slip
24. Surnom de l'année 2015
25. Ce qu'il peut être malicieux
26. AKA Gariguette, AKA Plougastel
27. Une varrrrrriante du pot au feu
28. On est pas dans le 16ème
29. Nantes est en Bretagne
30. La dernière des fill'd'Sedan
31. Z'est zlave ?
32. Reste-t-il du Chardonnay, d'ailleurs ?
33. AKA Chef Cadeau CIEM (CC CIEM)
34. Une de ces quatre qualités est très rare à Paris
35. #TigreTresFeroce

les commentaires ont toujours été utiles.

Un énorme merci également au labo halieutique de l'Agrocampus chez qui j'ai pu passer des supers moments. Un merci spécial à Catherine, qui m'a aidé comme elle a pu devant mon handicap administratif [36], et à qui j'ai failli coûter la vie lors d'une danse bordelaise. Un grand merci aux thésards locaux qui m'ont enseigné l'halieutisme et ses principes fondamentaux. Je parle évidemment de Benoit, Féfé, Schnappy et Émilie ! Les quelques verres que nous avons pu partager m'ont donné un goût de trop peu ! Un petit coucou à Maxou [37] et Fab, qui feront une grande carrière dans la recherche, c'est certain. Un grand merci à Jérôme pour tes leçons de bases de données ! Merci également à Olivier, Didier et Hervé pour leur sympathie malgré mon ignorance en science halieutique. Sachez que j'ai appris beaucoup de choses en vous écoutant !

Un grand merci également aux différents membres du groupe trajectométrie, je pense en particulier à Nico et Sophie avec qui j'ai passé des très chouettes moments. Également un grand merci à mon collègue stagiaire, Guillaume [38], Monsieur Éléphant de Mer, pour tous les bons moments passés ensemble.

Heureusement, en dehors des harassantes heures de labeur, il y avait les "gens extérieurs", mes colocataires. Un grand merci à Xanxandre, qui m'a accueilli quelques jours lors de mon arrivée à Nantes. Merci à toi de m'avoir aider à penser différemment, merci pour les deux ou trois soirées que nous avons pu faire. Merci pour les fous rires sur toutes ces situations absurdes. Merci pour m'avoir empêché de donner un sens à ces 4 années à vivre avec toi. Tu as été le roi de la jungle [39] de Bouffay.
Un grand merci à mon autre colocataire, Bobo, Stook, Rodrigue, Merguez, Michal [40], l'Homme aux 100 surnoms. Merci pour ton humour ravageur, dévastateur, caustique, animal, Jean Pascal. Merci pour avoir été un phare dans cette grande tempête, un point fixe inamovible dans ce grand monde en mouvement qu'est la Science, repère absolu dans un univers beaucoup trop souvent relatif. Les reflets roux de ta barbe d'automne berceront pour toujours mon cœur d'une langueur monotone.
Grâce à vous deux, le Marlowe a passé une très bonne thèse, et ça, ça a beaucoup compté pour moi.

Et puis, sans ordre de préférence, un grand merci aux autres amis.
Les potes de la pilate, Titi, Eddy, Kev, Steph, Pupuce, Fanch, Yann et aux copines, Marine, Fanny, Valou, Mag, Gaëlle...
À Cisco, compagnon d'aventures, merci pour tous les moments passés ensembles avant, et pendant cette thèse, à Céline, pour ta bonne humeur et avoir réussi à ordonner le bonhomme.

36. Au fait, je suis inscrit ?
37. Bwrrrffff
38. À quelle heure est ton train ?
39. Un lion en Cage, en quelque sorte
40. Mais qui a volé ... ?

À Gaëlle, Doudou, Laurette, qui m'ont soutenu, de loin, mais soutenu quand même !
À Conseil, avec qui j'ai tellement vécu !
À Cart, Pitchou, Brieuc et Sussu pour tous les moments à Nantes et ailleurs pendant cette thèse !
À la team du Master, Marc, Antoine, merci pour les bons moments avant cette thèse, pendant, et j'en suis sûr, après.
Aux Montpellierains, Askia [41], pour ton accueil à Montpel, capitale du sud, et Bebel [42], pour ton soutien éternel au Rock store [43].
À la team Parisianno Montréalaise, Bastos, Léo, Manon, je garde l'œil sur vous !
Aux potes de la term, Simon et Camille, Jean, Namik, pour toutes les chouettes conversations, qui m'ont beaucoup appris...
Aux amis de Rennes, la team Normando Rennaise, Renaud, pour toutes ces discussions sur les maths, la vie et la poésie, Lolo [44], Jumpy, Clémence, Ketsia, Roomy, Pierrot, Chloé, Adrien, Marco avec qui j'ai passé des sacrés moments [45]
Finally, thanks to all my friends that I've met in Seattle. Thanks to Kotaro, who welcomed me when I arrived. And huge thanks to the biohouse, the greatest roommates in the greatest country on earth ! Thanks "Ze Boss" Daryl, Elisha, Marie, "The Grinch" Peter, and "Doodle Doo" Frazier ! Hope to see you again guys for some wine and cheese party [46]

Je tiens à faire un merci spécial à mes 3 frères, Yannick, pour ton insatiable curiosité, ton amour de la connaissance, et toutes les discussions que nous avons pu avoir. Léo, pour ton charisme, ta force, tes récits de voyages et les béquilles que tu m'as données durant toute ma jeunesse. Gilles pour tes passages à Nantes absurdes, ton vin de je ne sais quoi, tes légumes saumurés, ton kéfir. Bref tous ces trucs là ! Merci à vous les frangins !
Un grand merci également à Aurélie, Ethan, et Liam ! Sacrée famille !
Enfin un grand merci à Papa et Maman, sans qui rien de tout ça n'aurait eu lieu.

41. Vous l'avez lu ce livre ?
42. L'As des As, le Professionnel
43. Bip-Bip !
44. MIPE-MIPE !
45. Andalousie, je me souviens...
46. With butter

Table des matières

Liste des figures

Liste des tableaux

And I know I will be loosened
From the bonds that hold me fast
And the chains all around me
Will fall away at last

And on that grand and fateful day
I will take thee in my hand
I will ride on a train
I will be the fisherman

With light in my head
You in my arms.

The Waterboys, Fisherman's Blues

Chapitre 1

Introduction

1.1 Contexte

Le mouvement, aspect fondamental des sciences physiques et de la vie, est depuis longtemps formalisé et étudié dans différents contextes. Le mouvement a ainsi très tôt donné lieu à des protocoles d'observations permettant l'élaboration de différents modèles.
En astronomie, les différents cycles des astres autour de la Terre furent précisément observés par les Grecs. Ces observations servirent de base à un modèle formalisé par Ptolemée dans son Almageste, décrivant ainsi les mouvements du Soleil et de la Lune autour d'une Terre immobile. De la même manière les observations de Copernic, Kepler et Galilée viendront alimenter l'étude approfondie du mouvement des corps célestes, et de sa formalisation mathématique, par Newton. Dans un schéma similaire, on peut citer le botaniste Robert Brown, dont les observations du mouvement de particules composant les grains de pollen (Brown, 1828) donnèrent lieu à la formalisation mathématique du mouvement Brownien par Einstein (Einstein, 1906).

Dans le domaine du vivant, l'étude du mouvement des animaux n'a pu, historiquement, se faire de manière aussi systématique que celles des astres. Si Aristote, au IVe siècle avant notre ère, traite déjà "Du mouvement des animaux", décrivant ainsi les moteurs du déplacement animal, ce n'est que plus récemment que des modèles formalisés ont été proposés. Prenant inspiration dans les modèles utilisés pour les particules, tels que le mouvement Brownien, des modèles théoriques se basant sur un formalisme mathématique ont alors été proposés pour le mouvement animal (Skellam, 1951). Le mouvement est alors devenu part intégrante de la théorisation des stratégies animales (Schoener, 1971; Kiester and Slatkin, 1974). Les progrès technologiques ont permis l'acquisition, depuis le début des années 1990, de vastes banques de données sur le déplacement des organismes vivants [1]. L'utilisation de marqueurs radios, acoustiques, GPS, ont permis l'observation de nombreuses trajectoires d'organismes de divers rangs taxonomiques, allant du bar (*Dicentrarchus labrax*, projet BARGIP [2]), aux éléphants de mer (*Mirounga leonina*, Guinet *et al.*, 2014), en passant par les cigognes blanches (*Ciconia*

1. Dont une est en libre accès : `www.movebank.org`
2. `http://wwz.ifremer.fr/bar`

ciconia[3]). Ces données ont ouvert de nouvelles perspectives pour l'écologie animale (Cagnacci *et al.*, 2010). Elles permettent une nouvelle analyse du comportement de recherche alimentaire, comme le calcul de nouveaux indices pour la qualité de prédation (Merrill *et al.*, 2010) ou de tester la validité de modèles théoriques comme sur la notion de "optimal foraging" [4] chez les herbivores (Owen-Smith *et al.*, 2010). De plus, les données GPS offrent de nouveaux moyens pour définir l'espace vital d'un individu (Kie *et al.*, 2010), permettent de mieux définir le lien entre mouvement individuel et dynamique des populations (Turchin, 1998; Morales *et al.*, 2010), ou d'évaluer la réponse de certains animaux aux facteurs anthropiques (Preisler *et al.*, 2013).

D'autre part, dans de nombreuses sciences humaines et sociales, l'observation du mouvement a servi à la mise en relief de schémas récurrents. En sciences de l'information géographique, des données de déplacement de voyageurs sont utilisées pour élaborer des schémas d'accessibilité (Kwan, 1998), des données de téléphones portables sont utilisées pour décrypter le déplacement de piétons dans des foules, dans les supermarchés, ou dans les villes (voir Long and Nelson, 2013 pour les références en "time geography"). En urbanisme, des données de déplacement des véhicules sont utilisées pour élaborer des modèles de comportement des conducteurs (Ding *et al.*, 2015). De la même manière, des données de mouvement de l'œil sont utilisées en neurologie (Rutishauser and Koch, 2007), en traitement d'image (Liu and Heynderickx, 2009), ou encore en psychologie (Stacey *et al.*, 2005) pour valider différentes théories dans ces disciplines.
L'Homme étant lui même un grand prédateur, son activité dans l'écosystème est aussi l'objet de l'étude du mouvement. Ainsi, en halieutique [5], pour des problématiques de gestion et de contrôle, de nombreux programmes de suivi des navires ont été mis en place, utilisant les Vessel Monitoring System (VMS). Dans une littérature récente, ces données furent analysées en utilisant des modèles développés pour les grands prédateurs (Bertrand *et al.*, 2007), et ont permis de caractériser, à petite échelle, l'activité de pêche des navires européens.

Dans tous ces domaines, l'analyse du mouvement est faite dans un cadre quantitatif, tel que préconise Turchin dans son ouvrage de référence *Quantitative analysis of movement* :

> *« [The fact] that ecology is, and should be, a quantitative science becomes especially clear when we begin considering issues that are relevant to the society whithin which we live – the same society that provides funds for our research. »*
>
> **Turchin, 1998**

Cette approche quantitative doit permettre de décrire et de quantifier les phénomènes afin d'aborder des questions concrètes pour la gestion (où, quand, combien ?). De par la formalisation du vivant, un enjeu important est également de pouvoir prédire celui-ci (Evans *et al.*, 2012). Cette formalisation passe par le langage mathématique, outil essentiel en écologie, comme en physique, pour décrire le mouvement.

3. `www.bto.org/science/migration/tracking-studies/stork-tracking/`
4. Quand aucune traduction générique n'a été trouvée, le terme original est conservé
5. Science de l'exploitation des ressources aquatiques vivantes

Dans les exemples énumérés ci dessus, on peut distinguer deux approches pour l'étude du mouvement.

1. Une approche descriptive : Le mouvement est alors l'objet d'étude en lui même. Son observation permet de confirmer ou de contredire des hypothèses :
   - La Terre est immobile ;
   - Les routes migratoires de tel animal passent par tel endroit ;
   - Les navires de pêche ne volent pas [6] ;

En écologie, par exemple, cette approche est souvent utilisée pour évaluer l'importance de facteurs environnementaux. La significativité de leur influence sur les trajectoires observées est testée à l'aide de modèles statistiques (Avgar *et al.*, 2013, par exemple). Un principal frein à cette approche est la nécessité de mesurer, à une échelle en adéquation avec celle de la trajectoire, les facteurs explicatifs.

2. Une approche mécaniste ; Le mouvement est perçu comme la conséquence d'un (ou de) phénomène(s) d'intérêt non observé(s). Son observation doit permettre d'expliciter ce(s) mécanisme(s) sous-jacent(s) :
   - La force de la gravitation ;
   - La présence de nourriture, différents comportements adoptés (chasse, besoin de reproduction) ;
   - Des intérêts économiques, divers comportements sociaux.

Cette deuxième approche mécaniste nécessite une formalisation de la relation Phénomène (Cause)-Mouvement (Conséquence). Nous appellerons cette formalisation un modèle de mouvement.

1.2 Une approche mécaniste du mouvement

Turchin définit le mouvement comme le "processus par lequel les organismes individuels sont déplacés dans l'espace, au cours du temps" (Turchin, 1998). Dans leur article sur l'édification d'un paradigme pour l'étude du mouvement, Nathan *et al.* (2008) définissent celui-ci comme un "changement de position spatiale d'un individu au cours du temps". Dans la suite, nous préférerons la définition de Turchin, à une distinction près, nous considérons le mouvement comme le "processus par lequel les organismes individuels se déplacent dans l'espace, au cours du temps". Cette nuance traduit le fait que les modèles développés par la suite ont vocation à décrire le mouvement d'organismes actifs dans leur déplacement (excluant les déplacements exlusivement passifs, comme celui du pollen dans le vent, du plancton dans les courants océaniques...).

1.2.1 Trajectoire réelle et trajectoire observée

On définit une trajectoire comme la réalisation du mouvement, et donc comme l'ensemble des couples (position, date d'acquisition) pris par un individu au cours d'une fenêtre de temps.

6. Maman, les p'tits bateaux qui vont sur l'eau n'ont donc pas d'aile

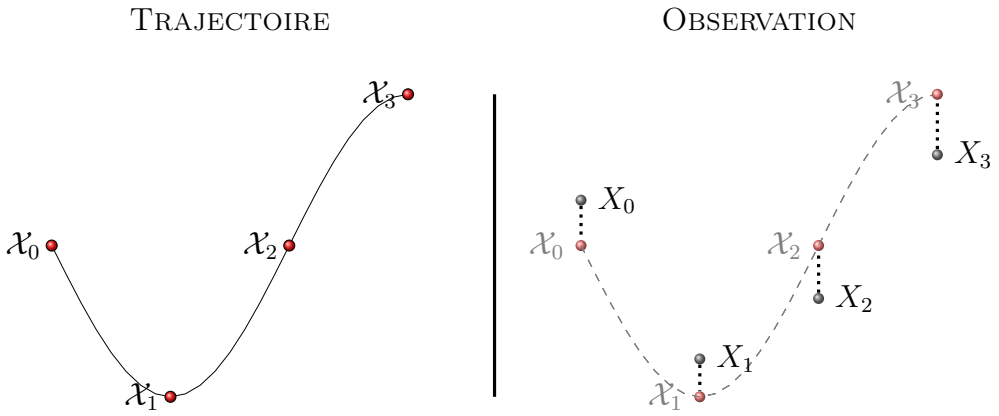


Figure *1.1 – Différence entre la trajectoire (à gauche) et son observation (à droite). Le processus de gauche est continu. L'observation de droite est discrète. L'écart entre la position observée et la vraie position représente une potentielle erreur de mesure.*

Formellement, on définit une trajectoire comme un processus $(\mathcal{X}_t)_{0\leq t\leq T}$ [7]. Les positions prennent valeur dans un ensemble $\mathcal{E}$ (inclus dans $\mathbb{R}^2$) [8]. Les temps associés à ces positions prennent valeur dans un intervalle $[0, T]$, indiquant les temps de début et de fin de l'observation. Il est nécessaire de faire une distinction fondamentale entre la trajectoire $(\mathcal{X}_t)_{0\leq t\leq T}$ et son observation, pour deux raisons :

- Une trajectoire continue est observée de manière nécessairement discrète [9] aux temps $t_0 = 0, \ldots, t_n = T$;
- Dans le cas général, en raison de multiples sources d'erreurs de mesure, la position enregistrée à l'instant t peut être différente de la vraie position.

Ainsi, une trajectoire observée sera une série de positions $X_0, \ldots, X_n$ acquises aux temps $t_0 = 0, \ldots, t_n = T$, (figure 1.1). Dans le cas où l'observation est faite sans erreur (la position observée est la vraie position), on aura donc $\mathcal{X}_i = X_i$. Pour les modèles développés dans la suite, cette hypothèse sera faite (sauf mention contraire).

1.2.2 Une modélisation individuelle du déplacement

Dans son livre, Turchin (1998) a pour objectif de "mesurer et modéliser la redistribution des populations chez les animaux et les plantes" [10]. L'auteur fait la distinction entre deux cadres pour modéliser le mouvement. Il s'agit d'un cadre "population", utilisé généralement pour la dispersion des microorganismes et d'un cadre "individuel", plutôt utilisé pour le déplacement de gros organismes. Cette distinction trouve son origine en mécanique, dans l'étude de la vitesse des particules composant un fluide, qui peut s'envisager soit dans le cadre Eulérien ou dans le cadre Lagrangien :

7. Dans la suite, on considèrera que la première position est au temps 0
8. Les trajectoires considérées dans ce manuscrit sont dans l'espace en 2 dimensions
9. Même si les temps d'acquisition sont très fins, comme lors d'un suivi vidéo, la donnée reste discrétisée (à la fraction de seconde, mais discrétisée tout de même)
10. Traduction libre

Le cadre Eulérien (ou cadre "flux")
D'un premier point de vue, le plus développé par Turchin, l'observateur se place à un point fixe M de l'espace. On observe au cours du temps le passage d'un fluide au point M. Ainsi, les vitesses mesurées au cours du temps (le flux) sont celles d'un ensemble de particules composant le fluide, ensemble passant à un moment par le point M.

Le cadre Lagrangien (ou cadre "particule")
De ce point de vue, l'observateur est attaché à une particule (ou à un ensemble de particules) composant le fluide au cours du temps. Ainsi la vitesse mesurée est celle de la (ou du groupe de) particule(s) se déplaçant dans l'espace.

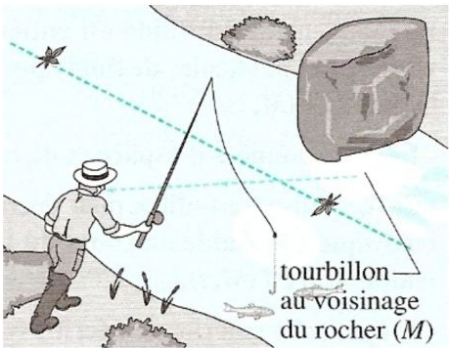
Un exemple de la différence des deux formalismes est montré sur la figure 1.2. Ces deux points de vue sont intimement liés par l'équation de Fokker Planck (Turchin, 1998). Pour étudier le mouvement des insectes et des microorganismes, l'approche Eulérienne reste l'approche majoritaire (et souvent la seule possible). En effet, dans ce cas, le suivi d'un individu est matériellement difficile. De plus, le suivi individuel n'a pas nécessairement d'intérêt, quand un comportement moyen va permettre de répondre aux questions posées. De manière plus générale, cette approche est très utilisée pour les modèles d'habitats, ou de niches. On s'intéressera alors à la densité d'une population ou à la probabilité de présence d'un individu dans un lieu donné :

> *« Eulerian approaches (...) can be used to predict [expected patterns of space use], because they focus on how the probability of an individual's occurrence (or, if studying a population, the density of animals) can be expected to change through time at any given point in space. The models thus become "place-based". »*
> **Smouse *et al.*, 2010**

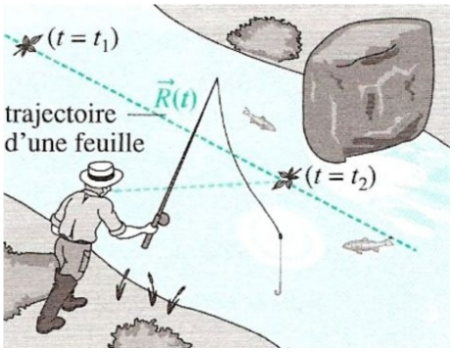
D'un autre côté, pour des organismes dont le suivi individuel (type GPS) est disponible, l'approche Lagrangienne permet de répondre à d'autres questions. En s'intéressant à la trajectoire de l'individu, on s'intéresse directement au processus adopté par l'individu (ou un groupe d'individus) au travers de son environnement (Nathan *et al.*, 2008; Smouse *et al.*, 2010). Cette approche est particulièrement adaptée pour détecter les différents comportements d'individus au cours de leur déplacement (Smouse *et al.*, 2010). De plus, l'approche Lagrangienne permet de définir des modèles "agent" spatialement explicites (Tang and Bennett, 2010).

1.2.3 Les mécanismes du mouvement

Une vision mécaniste du mouvement postule que les trajectoires d'un individu sont les conséquences de mécanismes sous-jacents non observés. Ces mécanismes sont ceux que les écologues veulent déterminer et comprendre.
Un cadre général à cette approche est proposé par Nathan *et al.* (2008), qui décrit le mouvement comme la conséquence de quatre blocs fondamentaux. Les trois premiers sont des réponses à différentes questions "Pourquoi ?", "Comment ?", "Où et quand ?".



(a) Formalisme Eulérien, le pêcheur se concentre sur un point fixe du fond de la rivière (le point M)



(b) Formalisme Lagrangien, le pêcheur se concentre sur un point mobile de la rivière (la feuille)

FIGURE *1.2 – Un pêcheur évalue le courant dans une rivière. Différence entre les formalismes Eulérien et Lagrangien présentés en section 1.2.2 (source* `http://olivier.granier.free.fr`*)*

1. Pourquoi bouger ? L'individu présente un état interne (internal state) décrivant sa physiologie et sa psychologie (faim, phase de reproduction...) ;

2. Comment bouger ? L'individu présente une capacité motrice (motion capacity) décrivant ses capacités biomécaniques à se déplacer (capacités de vol, de nage...) ;

3. Où et quand bouger ? L'individu présente une capacité à la navigation (navigation capacity) décrivant ses capacités à orienter son mouvement et à le planifier dans le temps (capacité de perception, mémoire ...)

Ces trois blocs fondamentaux interagissent entre eux, et avec un quatrième compartiment :

4. Quelle influence de l'environnement ? L'individu est influencé par des facteurs externes (external factors), qui modifient son déplacement.

Le déplacement observé est donc la conséquence de l'interaction de ces quatre blocs, qu'il convient de bien identifier pour chaque individu. En pratique, il est certainement difficile, dans la plupart des cas, de bien séparer ces quatre blocs. Cependant, cette vision a l'avantage d'unifier en un seul paradigme les approches mécanistes du mouvement.
On ne cherche donc pas ici à faire du test d'hypothèse sur ce qui explique le mouvement, toute chose étant observée par ailleurs (ce que nous avons appelé l'approche descriptive dans la section 1.1). En effet, la mesure des différents facteurs, internes (comportementaux) ou externes (environnementaux), explicatifs du mouvement est, en pratique, souvent impossible. Par une approche mécaniste, on considère ainsi le mouvement comme un potentiel révélateur de ces facteurs.

Cette vision mécaniste de Nathan *et al.* (2008) appelle à une modélisation paramétrique du mouvement. Comme dit en section 1.1, le modèle est le résultat de la formalisation de la relation Phénomène (Cause)-Mouvement (Conséquence). Dans le cadre mathématique, on considère que le mouvement est quantifiable, il faut donc explicitement le quantifier.

1.2.4 Un temps continu ou discret ?

Nous avons vu en section 1.2.1 qu'il convenait de distinguer la trajectoire de son observation. Entre ces deux niveaux doit se placer le modèle de mouvement formalisé. On a donc trois niveaux, dont deux sont concrets, et l'autre est à formaliser :

	Objet	Environnement temporel
1)	Trajectoire	Continu
2)	Modèle de mouvement	Discret ou continu ?
3)	Observations	Discret

Il a été montré que le choix discret/continu n'était pas anodin, tant dans la formalisation du modèle, que dans les résultats obtenus sur l'application aux données (McClintock *et al.*, 2014). Le choix du cadre de formulation du temps et l'analyse des conséquences de ce choix sont donc fondamentaux. Comment choisir si le modèle doit être continu ou discret en temps ? Dans une revue récente de modèles de mouvement, McClintock *et al.* (2014) expriment ce point de vue :

> *« In our experience, the greatest source of confusion among practitioners [using movement models], both in terms of implementation and biological interpretation, seems to be the distinction between continuous- and discrete-time formulations of the movement process. »*
>
> **McClintock *et al.*, 2014**

Selon les auteurs, la modélisation discrète serait plus naturelle pour les utilisateurs. Cette vision a, de notre point de vue, deux origines. La première est que la nature des observations conditionne souvent le choix du modèle, d'où l'idée "naturelle" d'adopter un formalisme en temps discret.

De plus, un formalisme discret découle de l'idée intuitive qu'une trajectoire est une succession de pas ("steps", Turchin, 1998), définie par une longueur et une direction. Ces pas sont parfois appelés éléments fondamentaux du mouvement ("fundamental movement elements", Getz and Saltz, 2008). Cette modélisation discrète du mouvement est, selon McClintock *et al.* (2014), plus naturelle et mieux comprise par les utilisateurs, qu'une approche continue.

Dans un telle décomposition discrète, il convient pour l'utilisateur de choisir une unité de temps fondamentale. Ce choix doit se faire à une échelle adéquate au vu des phénomènes que l'on veut observer. Or cette adéquation suppose d'avoir la qualité d'échantillonnage nécessaire, ainsi qu'une bonne connaissance *a priori* des phénomènes. Le choix de cette unité temporelle est ainsi critique et influence les résultats (Codling and Hill, 2005; Fryxell *et al.*, 2008, entre autres).

D'un autre point de vue, on peut considérer le mouvement comme un processus continu. Le temps continu induit une plus grande complexité dans la formalisation mathématique du processus. L'absence de pas de temps minimal implique une écriture infinitésimale du modèle de mouvement qui peut être moins intuitive, ou moins facile à assimiler pour les utilisateurs.

Une autre distinction majeure entre modèles discrets et continus est la dépendance à l'échantillonnage. De par le besoin d'une unité de temps fondamentale, la majorité des modèles discrets repose sur une série d'observations espacées régulièrement, selon l'unité fondamentale de temps. Dans le cas contraire, les hypothèses du modèle de déplacement ne sont plus directement en phase avec les observations. Or, en télémétrie il n'est pas rare que l'acquisition des données soit irrégulière. Dans ce cas, pour se ramener au cadre du modèle, un choix important est à faire de la part de l'utilisateur concernant le pré-traitement des données (interpolation, suppression de points...). Cette contrainte n'a plus lieu d'être dans un modèle continu, où les mécanismes de déplacement sont décrits pour tout intervalle de temps. L'un des apports majeurs du formalisme continu est de ne pas être incompatible avec une réalité de l'échantillonnage : son irrégularité.

1.3 Une modélisation stochastique du déplacement

Notre étude a pour but de définir un modèle mécaniste de mouvement, afin de révéler des phénomènes d'intérêt écologique. Jusqu'à présent, cette démarche peut se rapprocher de celle de Newton, utilisant la chute d'une pomme pour révéler la force gravitationnelle. Cependant, en supposant qu'une mécanique déterministe existe dans le déplacement d'un individu (i.e. toutes choses étant égales par ailleurs, un individu effectuerait toujours le même déplacement), une description exhaustive est inaccessible. En effet, si l'on reprend par exemple le quatrième bloc du formalisme de Nathan *et al.* (2008), il est impossible, hors laboratoire, de mesurer continûment, en tout point de l'espace, toutes les conditions environnementales. Il en est d'ailleurs de même pour l'état interne de l'individu.
Cependant, pour le modélisateur, il est possible qu'une partie (connue ou non) de ce déterminisme soit non centrale pour la question posée. Dans ce cas, il peut être commode d'assimiler les variations (non mesurées) de ce déterminisme, à une variabilité stochastique. Le modèle décrit alors le mouvement comme la conséquence d'un phénomène que l'on veut expliciter, auquel vient s'ajouter une stochasticité. Le cadre stochastique, basé sur le formalisme des probabilités, doit permettre d'estimer les phénomènes d'intérêt pour les écologues. Il ne se veut pas être la traduction d'une propriété intrinsèquement aléatoire du déplacement animal. On se réfèrera ici à Georges Matheron, discutant de l'objectivité des modèles probabilistes :

> *« [L]'"aléatoire" n'est en aucune façon une propriété, univoquement définie, ni même définissable, du phénomène lui même. Mais, uniquement, une caractéristique du, ou des, modèles que nous choisissons pour le décrire, l'interpréter, et résoudre tel ou tel problème que nous nous posons à son sujet. »*
> **Matheron (1978), Estimer et Choisir, Introduction**

Le choix du stochastique peut donc être vu comme un choix de simplification de la réalité. Ce choix est particulièrement adapté quand le but est d'expliciter des comportements au cours de la trajectoire, par exemple. Dans ce cas, le but n'est pas de reproduire exactement les propriétés de la trajectoire d'un individu (passage systématique par un point précis, par exemple), là où un modèle stochastique aurait du mal à reproduire les schémas observés.

1.3.1 Un formalisme stochastique

Dans un modèle déterministe de mouvement (comme ceux employés en mécanique classique par exemple) dépendant de paramètres θ, la trajectoire d'un individu, étant données des conditions initiales, est une réalisation unique d'une fonction de θ et des conditions initiales.
Cependant, cette formulation ne peut avoir de réalité pratique en écologie. En effet, comme dit plus haut, l'ensemble des facteurs déterminant le déplacement étant impossible à mesurer exhaustivement, l'ensemble θ des paramètres est alors nécessairement incomplet. Il est ainsi plus commode de considérer un processus stochastique à trajectoires continues [11], tel que chaque trajectoire [12] est une réalisation de ce processus. Une trajectoire est alors la réalisation d'une variable aléatoire sur un espace de dimension infinie, l'espace des fonctions continues dans $\mathbb{R}^2$.
Dans la pratique, un modèle de mouvement paramétré par θ consiste à décrire la loi de transition du processus, qui spécifient la loi de la variable aléatoire $\mathcal{X}_{T+\Delta}$ (en 2 dimensions) connaissant le processus $(\mathcal{X}_t)_{0\leq t\leq T}$, et le paramètre θ.
Un modèle simple utilisant ce formalisme est la marche aléatoire Gaussienne dans $\mathbb{R}^2$, de paramètre σ^2. Dans ce cas, la loi de $\mathcal{X}_{T+\Delta}$ connaissant $(\mathcal{X}_t)_{0\leq t\leq T}$ celle d'une loi Gaussienne centrée en $\mathcal{X}_T$ et de variance $\Delta\sigma^2$ (pour les modèles de marches aléatoires développés en écologie animale, on peut se référer à Codling *et al.*, 2008, par exemple). Dans ce modèle, la connaissance de tout le passé se résume à sa dernière information, c'est un modèle Markovien.

À un modèle de mouvement, on peut ajouter une deuxième couche stochastique, qui est un modèle d'observation, quand l'observation X_t est différente de la réalisation $\mathcal{X}_t$ de la trajectoire au temps t. Dans ce manuscrit, comme dit en section 1.2.1, nous resterons le plus souvent dans le cas sans erreur, c'est à dire $\mathcal{X}_t = X_t$.

1.3.2 L'estimation des paramètres

Les modèles stochastiques de mouvement proposés en section 1.3.1 sont donc dépendants d'un ensemble de paramètres θ. Ces paramètres sont la formalisation de phénomènes que l'on veut estimer à partir des observations. La mise en place du modèle doit être suivie par la mise en place d'une technique d'estimation des paramètres.
Dans ce travail, nous nous placerons dans le cadre d'estimation par maximum de vraisemblance. Si on considère un ensemble d'observations de trajectoires (noté $\mathbf{X}$), et un jeu de paramètres θ on peut, en théorie, calculer la densité de probabilité $g(\mathbf{X}|\theta)$ des observations pour le modèle. Cette fonction étant calculable pour tout θ, on peut définir la fonction de vraisemblance :

$$L(\theta) = g(\mathbf{X}, \theta) \tag{1.1}$$

11. Soit $\mathbb{T} \subset \mathbb{R}_+$. On appelle processus stochastique une famille de variables aléatoires définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé, et indexées par le temps $\mathcal{X} = \{\mathcal{X}_t, t \in \mathbb{T}\}$

12. Une trajectoire d'un processus stochastique est une fonction $t \in \mathbb{T} \mapsto X_t(\omega)$ pour $\omega \in \Omega$ fixé.
Lorsque ces fonctions sont continues, on peut donc aussi considérer qu'un processus stochastique $\mathcal{X}$ est une variable aléatoire à valeur dans l'espace $\mathcal{C}(\mathbb{T}, \mathbb{R}^2)$

L'ensemble de paramètres choisi sera alors l'estimateur

$$\hat{\theta} = \text{argmax } L(\theta) \tag{1.2}$$

Il est fréquent, dans le cadre des modèles de mouvement, que la fonction de l'équation (1.1) ne soit pas calculable directement. Dans ce cas, il faut utiliser une méthode d'estimation adaptée pour estimer $\hat{\theta}$. Les méthodes d'estimation sont grandement dépendantes du modèle. De manière générale, leur complexité est variable. Par exemple, les modèles continus discutés en section 1.2.4 demandent un formalisme plus sophistiqué, ce qui explique en partie leur manque de popularité actuel (McClintock *et al.*, 2014). Les méthodes d'estimation choisies dans ce travail seront détaillées dans le corps des chapitres. Le choix de la méthode d'estimation n'est pas unique, il sera donc discuté en cours de chapitres, et plus largement abordé en discussion.

1.4 Les modèles de mouvement pour l'écologie animale

Le cadre de modélisation présenté ci-dessus est aujourd'hui très utilisé en écologie animale. Le modèle longtemps privilégié a été celui de la marche aléatoire (popularisé par Pearson en 1905), qui fut l'objet de nombreuses extensions (Turchin, 1998; Gurarie, 2008, pour un historique très intéressant). L'utilisation de ces modèles permet de mettre en évidence un caractère diffus (marche aléatoire "pure") ou au contraire, un caractère dirigé pour les trajectoires (marche aléatoire corrélée Codling *et al.*, 2008; Schick *et al.*, 2008).

1.4.1 Le mouvement pour connaître le comportement

Cependant, ces modèles, traduisant un unique type de mouvement le long d'une trajectoire, peuvent s'avérer insuffisants (Morales *et al.*, 2004). En effet, au cours d'un intervalle de temps, un animal peut adopter plusieurs comportements, du fait de l'environnement, ou de son état interne. Le mouvement de l'individu peut alors s'en ressentir. Une question majeure en écologie animale est de détecter différents comportements adoptés au cours d'une trajectoire (Firle *et al.*, 1998; Morales and Ellner, 2002). Certaines approches descriptives, utilisant des outils de mesure de l'état interne de l'animal ont été utilisées dans ce but (sur les saumons notamment, Hinch *et al.*, 2002). Comme évoqué plus haut, ces multiples mesures ne sont pas toujours disponibles. Dans ce cas, des approches mécanistes ont été proposées pour déceler différents comportements. Le principe de base de ces approches est que la trajectoire est la somme de plusieurs morceaux homogènes [13], et que chacun de ces morceaux est le reflet d'un comportement. La trajectoire doit ainsi permettre de retracer la séquence de ces comportements.
Pour atteindre cet objectif, la grande majorité des études se base sur un formalisme en temps discret. Comme dit plus haut, on considère alors que le mouvement est constitué de pas (ou éléments) fondamentaux (Turchin, 1998; Getz and Saltz, 2008). Si ces éléments ont une réalité

13. Ce que l'on entend ici par homogène sera explicité dans le chapitre 2

physique [14], ce n'est pas nécessairement le cas d'un comportement. Celui-ci doit être défini *a priori* par l'utilisateur, en fonction de l'échelle d'observation :

> *« [T]o appropriately characterize movement components of an individual's path over time, we should endeavor to identify canonical activity mode (CAM) distributions that emerge from the mix of fundamental movement elements (FMEs) that characterize the activity in question : i.e., CAMs are composites of the FMEs, and their characteristic step size and direction of heading distributions will depend on the length of sampling intervals and scale of analysis. »*
>
> **Getz and Saltz (2008)**

La figure 1.3 propose une représentation de cette idée. Le schéma voulu ici est donc de choisir *a priori* des comportements qui se distinguent par la distribution des vitesses et des directions. Pour intégrer cette idée dans un formalisme mathématique, les modèles à espace d'états ont largement été utilisés (Jonsen *et al.*, 2003; Patterson *et al.*, 2008; Jonsen *et al.*, 2013a). Les modèles à espace d'états permettent de considérer la trajectoire observée comme la couche inférieure d'un modèle étagé. L'étage d'intérêt est alors celui des comportements. Il est caché (car non observé), et est estimé grâce aux données. Une présentation plus détaillée de ces modèles sera faite dans le chapitre 2.

De nombreuses applications de ces modèles ont été proposées en écologie animale [15], sur des mammifères terrestres (Morales *et al.*, 2004; Langrock *et al.*, 2012), des mammifères marins (Jonsen *et al.*, 2005; Breed *et al.*, 2012; McClintock *et al.*, 2012), des reptiles (Jonsen *et al.*, 2006, 2007), ou encore des oiseaux (Roberts *et al.*, 2004; Guilford *et al.*, 2004; Freeman *et al.*, 2013). Ces études utilisent toutes un formalisme mécaniste et Lagrangien pour établir des schémas comportementaux des espèces étudiées. Les modèles à espace d'états peuvent aussi être utilisés dans un cadre Eulérien, toujours dans une problématique de détecter les comportements (Pedersen *et al.*, 2011, par exemple, pour les thons rouges).

1.4.2 Le mouvement : utilisation et perception de l'environnement

Beaucoup d'animaux sont non nomades et se déplacent dans un environnement restreint. De ce constat est né le concept de "home range" [16] (Burt, 1943), décrit originellement ainsi :

> *« [A]rea traversed by the 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 »*
>
> **Burt (1943)**

Cette définition intuitive reste non quantitative sur la manière de définir la zone et les sorties "occasionnelles" de celle-ci. L'utilisation des données de trajectoires a participé à la définition de méthodes statistiques d'estimation du "home range". Une méthode très répandue est celle du

14. Dans ce paradigme, en tous cas
15. La liste donnée n'est pas exhaustive
16. Parfois appelé domaine vital dans la suite

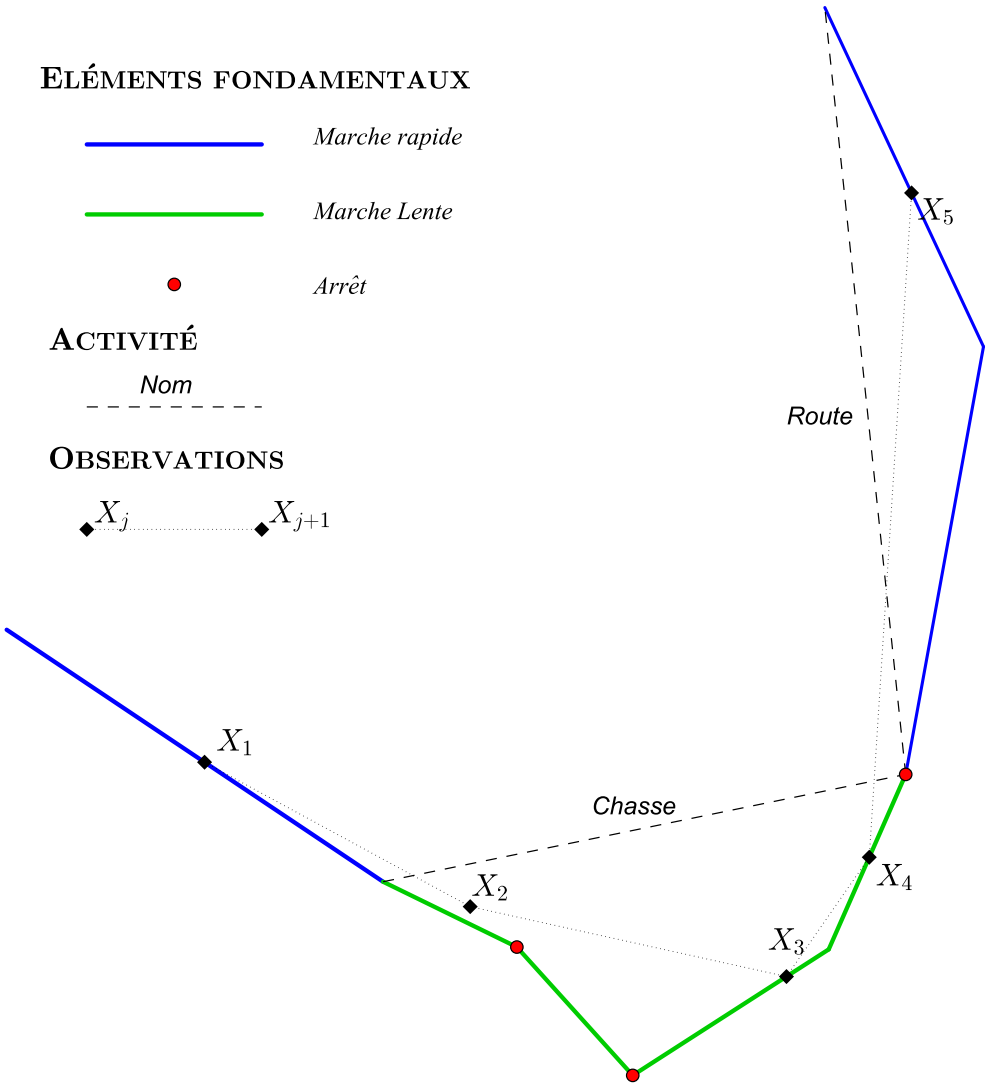


FIGURE *1.3 – Une trajectoire telle que formalisée par Getz et Saltz. Le mouvement est composé d'éléments fondamentaux. Ici la marche rapide, en bleu, la marche lente, en vert, l'arrêt, en rouge. Un comportement (ou activité) est défini a priori, et doit pouvoir être exprimé comme composé des éléments fondamentaux. Le but de l'estimation est, à partir des points d'échantillonnage, de distinguer ces comportements, ici la chasse et la route vers zone. Figure inspirée de l'article de Getz and Saltz (2008).*

polygone convexe minimal enveloppant les trajectoires observées (Hayne, 1949). Afin d'ajouter une notion d'intensité d'utilisation au concept de Burt, les écologues se sont concentrés ensuite sur la notion de densité d'utilisation (Utilization Distribution). Ce concept renvoie à la "distribution [de probabilité] des positions d'un individu dans le plan" (Worton, 1989). Cette densité est généralement estimée en utilisant des méthodes à noyau de densité (Laver and Kelly, 2008; Keating and Cherry, 2009), sur les trajectoires observées. Ces techniques font toujours l'objet de recherches actives (Fleming *et al.*, 2015).

À ces approches purement statistiques ce sont ajoutées des approches mécanistes. La vision mécaniste du "home range" se base sur l'étude du processus déterminant l'émergence du domaine vital (Kie *et al.*, 2010). En effet, ce dernier peut être vu comme une carte cognitive de l'animal, c'est-à-dire la perception spatiale de son environnement (Powell, 2000). Cette carte intègre la perception que l'individu aurait de la ressource, des zones de passages, des abris, de la présence de congénères. Dans une approche mécaniste, on suppose que cette carte contrôle le mouvement, donc que ce dernier révèle la carte. Ainsi, des modèles de mouvement ont été proposés liant directement le mouvement au domaine vital (ou à la densité d'utilisation). Ces modèles, en temps et en espace continu, font des hypothèses sur les équations du mouvement. Dans un cadre Lagrangien, le processus de Ornstein Ulhenbeck (ou un mélange de tels processus) a été proposé. Ce processus implique que la densité d'utilisation de l'individu est une loi Gaussienne (Dunn and Gipson, 1977), ou un mélange de lois Gaussiennes, mélange dû à de multiples comportements (Blackwell, 1997), ou à de multiples habitats (Harris and Blackwell, 2013). Le cadre Eulérien a également été proposé pour décrire la densité d'utilisation, dans un cadre multi habitats (Moorcroft *et al.*, 2006).

Une autre approche proposée place la carte estimée comme contrôle *a priori* de la trajectoire, et non plus comme un résumé *a posteriori* de celle-ci. Ce modèle décrit la trajectoire comme le reflet du potentiel de l'environnement (Preisler *et al.*, 2004; Brillinger, 2010; Preisler *et al.*, 2013). Dans cette approche, la finalité n'est pas d'estimer la densité d'utilisation de l'individu, mais de déterminer la manière dont celui-ci perçoit son environnement. On rejoint ici l'idée que le domaine vital est la conséquence d'une carte cognitive de l'individu (Powell, 2000).

1.5 La modélisation de trajectoires pour l'halieutique

L'effet de la pêche sur les écosystèmes est depuis longtemps reconnu et étudié. Dans une optique d'exploitation durable des ressources, l'organisation des nations unies pour l'alimentation et l'agriculture (en anglais, FAO) a prôné en 2001 [17] une approche écosystémique pour la gestion des pêches :

> *« L'approche écosystémique des pêches a pour objet de planifier, de valoriser et de gérer les pêches, en tenant compte de la multiplicité des aspirations et des*

17. Sommet de Reykjavik, `http://www.fao.org/docrep/005/y2198t/y2198t02.htm`

besoins sociaux actuels et sans remettre en cause les avantages que les générations futures doivent pouvoir tirer de l'ensemble des biens et services issus des écosystèmes marins »

FAO, 2001

L'Homme fait partie de l'écosystème, ainsi, son activité et ses impacts doivent être pris en compte dans cette approche. Le suivi des navires de pêche peut apporter une information sur plusieurs points clés de la gestion, par exemple :

- La cartographie de l'activité de pêche (Mills *et al.*, 2007) ;
- L'évaluation de l'impact des mesures de gestions sur la pêcherie (Hutton *et al.*, 2004) ;
- La distribution de la ressource ciblée (Walker *et al.*, 2015).

1.5.1 Le suivi individuel des navires de pêche : une nouvelle échelle

Dans cette optique, l'utilisation d'un système embarqué de suivi des navires de pêche, le Vessel Monitoring System (VMS), s'est imposée très tôt comme un outil de gestion des pêches. Commençant par les navires de plus de 24 mètres (1992) l'Union Européenne a imposé depuis 2005 (pour les plus de 15 mètres) et 2012 (plus de 12 mètres) [18] aux navires de pêche de s'équiper du VMS, (Kourti *et al.*, 2005).
La donnée VMS consiste en un GPS fixe embarqué sur les navires. Il s'agit donc d'une donnée non déclarative et automatisée. De par sa résolution temporelle (la base initiale réglementaire étant d'au moins une émission par tranche de deux heures [19], elle a été, en pratique, réduite à une position en moyenne toutes les heures pour les navires français), elle permet un suivi beaucoup plus fin de la distribution de l'activité que les données déclaratives de débarquements (ou données logbook).

Les données VMS fournissent donc, pour toute la flotte française de taille supérieure à 12 mètres, le suivi (plus ou moins) régulier des positions des navires à fine échelle. La marge d'erreur standard de positionnement du matériel VMS a été réduite à 10 mètres [20]. Si le but premier de cette base est le contrôle, elle est une base unique de mouvement pour le suivi des déplacements d'une population au cours du temps.
Le premier apport de ces données a été la vision à une nouvelle échelle de l'activité des navires de pêche. Ce changement d'échelle s'est effectué sur plusieurs points :

1. **Une échelle spatiale :** Les principales sources de cartographie de l'activité de pêche ont longtemps été les déclarations de débarquements (ou données logbook). Les zones de prise des captures débarquées y sont déclarées à l'échelle d'une zone réglementaire de plusieurs centaines de kilomètres carrés ($1 \times 0.5^\circ$). Les cartographies de l'activité de pêche en découlant ont donc une échelle minimale grossière au regard des zones observées (Pelletier and Ferraris, 2000; Hutton *et al.*, 2004). La donnée VMS donne une position

18. Règlement CE n° 1224/2009
19. Règlement CE n° 2244/2003
20. `http://www.fao.org/3/a-a0959e/`, page 41

ponctuelle des navires avec une erreur très faible, permettant ainsi un suivi plus fin des zones explorées.

2. **Une échelle temporelle :** Encore une fois, ce sont les données logbook qui renseignent le plus souvent les moments des captures. Ces données doivent décrire chaque activité, dans chaque zone (telle que définie ci-dessus), par tranche de 24 heures. Les données VMS fournissent un point sur la position des navires en moyenne toutes les heures.

3. **Une échelle agent :** La donnée VMS, de par sa finesse spatiale et temporelle a permis d'avoir une information importante à l'échelle du navire de pêche. Ce suivi permet des études individuelles (Bertrand *et al.*, 2007), qui apportent une information complémentaire à l'approche flottille (ou "population"), souvent adoptée en halieutique (Girardin *et al.*, 2015).

Cette nouvelle caractérisation fine de l'activité des navires de pêche permet ainsi de mieux estimer l'impact de la pression de pêche sur les écosystèmes (Poos and Rijnsdorp, 2007; Mills *et al.*, 2007), de mieux évaluer l'impact réel des mesures de gestion (Lehuta *et al.*, 2013), ou d'anticiper les réactions des pêcheurs à ces mesures (Vermard *et al.*, 2008).
À l'aide des données VMS, utilisant des réseaux de neurones, Russo *et al.* (2011b) ont identifié les métiers exercés par différents navires, à partir des caractéristiques de leurs trajectoires, passant ainsi de l'échelle individuelle à l'échelle flottille. Bertrand *et al.*, 2007, à l'aide de marche de Lévy, ont montré que les navires de la pêcherie d'anchois du Pérou avaientt des comportements similaires à ceux de certains oiseaux marins. Signe de l'utilisation croissante de ces données dans la recherche halieutique, une gamme d'outils pour l'analyse des données VMS est disponible pour le logiciel R (R Core Team, 2014) avec les packages **VMStools** (Hintzen *et al.*, 2012) et **VMSbase** (Russo *et al.*, 2014a).

1.5.2 Les données VMS pour déterminer l'activité de pêche

Comme dit plus haut, un intérêt majeur des données VMS est la détermination à une fine échelle des séquences d'activités de pêche. Cependant, les capteurs GPS ne donnent pas en temps réel l'information sur l'activité des navires. Différentes méthodes d'estimation des moments de pêche à partir des données VMS ont ainsi été développées récemment. La majorité des études se base directement sur la vitesse pour caractériser l'activité. Cette vitesse est, en général, la vitesse moyenne entre deux positions, calculée en faisant une interpolation linéaire pour la distance parcourue.

La méthode la plus répandue de détermination de l'activité consiste alors à définir un seuil en deçà duquel le navire est considéré comme en pêche (Palmer and Wigley, 2009). À partir des positions $X_0, \ldots, X_n$ observées aux temps $t_0, \ldots, t_n$, on calcule la série $V_0, \ldots, V_{n-1}$, telle que $V_i = \| (X_{i+1} - X_i) \| / (t_{i+1} - t_i)$. Le point X_i est alors dit "en pêche" si V_i est en dessous d'un certain seuil (4.5 noeuds, en France), en "non pêche" sinon.
Cette méthode de seuil peut également être couplée à la direction des navires (Mills *et al.*,

2007), améliorant ainsi sa performance. D'un autre côté, l'utilisation de réseaux de neurones a été proposée pour segmenter les trajectoires en pêche/non pêche, en se basant sur de multiples descripteurs des segments (heure, vitesse, accélération, changement d'angle) (Bertrand *et al.*, 2008; Joo *et al.*, 2011).

Plus récemment, en s'inspirant des modèles développés en écologie animale (section 1.4), l'estimation des moments de pêche à petite échelle s'est faite de manière plus mécaniste. Au travers de modèles à espace d'états, la trajectoire observée est décrite par un modèle de mouvement, dont les paramètres dépendent d'un état comportemental (pêche, route,...). La trajectoire observée est alors utilisée pour estimer conjointement les paramètres du mouvement ainsi que la séquence des comportements cachés. Ces méthodes ont été utilisées pour les chalutiers pélagiques du golfe de Gascogne (Vermard *et al.*, 2010), les thoniers seineurs de l'océan Indien (Walker and Bez, 2010), les chalutiers à crevettes et à coquilles saint Jacques d'Australie (Peel and Good, 2011), et les chalutiers pélagiques de la pêcherie d'anchois du Pérou (Joo *et al.*, 2013a). Ce dernier travail a d'ailleurs été l'occasion d'une comparaison de toutes les méthodes de détermination du comportement des navires de la pêcherie d'anchois du Pérou. Cette étude utilise des données d'observateurs embarqués comme validation des différentes techniques. Joo *et al.* (2013a) concluent que les meilleurs modèles sont les modèles de semi Markov caché, extensions des HMM. Ces modèles mécanistes l'emportent sur les méthodes de réseaux de neurones, de forêts aléatoires et de support vector machine (SVN).

L'utilisation de ces modèles mécanistes sur les données VMS est une illustration du paradigme souhaité par Nathan *et al.* (2008), décrit en section 1.2.3. Des hypothèses de mouvement sont analytiquement formulées et mises à profit pour l'estimation du comportement.
Cependant, dans les deux cas évoqués ici, ce n'est pas à proprement parler la trajectoire, mais un résumé de la trajectoire qui est utilisé pour détecter le comportement. Dans la méthode de seuil, la notion de séquence est même devenue inutile. Dans les modèle HMM, des mécanismes apparaissent, mais ils n'agissent que sur des quantités dérivées de la trajectoire [21].

1.5.3 Les données VMS pour définir les zones exploitées

Au travers des méthodes décrites dans le paragraphe précédent, les données VMS permettent une caractérisation spatiale de l'activité de pêche. Comme évoqué plus haut, ces estimations sont beaucoup plus fines que celles obtenues par les données déclaratives de débarquement, en plus d'être automatisées et non déclaratives. Cette nouvelle échelle s'avère nécessaire pour une gestion écosystémique des pêches (Gerritsen and Lordan, 2011).

Grâce aux données VMS, de nouvelles cartographies de l'activité de pêche ont donc été effectuées pour de nombreuses zones. Une telle cartographie à haute résolution a ainsi été réalisée pour la zone Ouest Irlande (Gerritsen and Lordan, 2011), et mise en relation avec les données de

21. Le fait de ne pas utiliser la trajectoire, mais un résumé de celle ci, sera discuté dans le chapitre 2

captures. Une cartographie similaire pour chaque engin de pêche a été publiée pour la Manche occidentale et le sud de la mer Celtique (Campbell *et al.*, 2014). L'utilisation de telles cartes a été proposée pour la gestion des pêches en Méditerranée (Martin *et al.*, 2014), notamment en étant intégrées à des modèles bio-économiques (Russo *et al.*, 2014b). Utilisant des données satellites de tous les navires (y compris ceux non contraints à l'équipement VMS), de telles cartes ont également été réalisées pour la flotte suédoise (Natale *et al.*, 2015). Tous les articles cités produisent donc des cartes d'activité de pêche (ou, après transformation, d'effort de pêche) estimée dans la zone d'intérêt.
Ces cartes sont toutes obtenues de manière analogue, à savoir :

1. Compter le nombre de points en pêche par une méthode de seuil de vitesse.
2. Les points estimés en pêche sont ensuite agrégés sur une grille de taille définie par l'utilisateur. La taille de cette grille étant dépendante de l'échelle d'observation des phénomènes.

Comme autre approche, le modèle mécaniste de comportement développé par Walker and Bez (2010) fut utilisé pour cartographier les zones d'activités de pêche, de recherche active et de route des thoniers seineurs au large de la Somalie (Bez *et al.*, 2011). Ce modèle hiérarchique s'appuie sur deux résumés de la trajectoire, les série des vitesses et des changements de direction (calculés). À un point est associé une vitesse et un changement de direction. À partir d'un HMM se basant sur ces deux caractéristiques, on assigne au point une activité (pêche, recherche active, ou route). La démarche est la suivante :

1. Pour chaque point de la trajectoire, une activité est assignée en utilisant le modèle HMM.
2. Pour chaque activité, les points estimés sont ensuite agrégés sur une grille de taille définie par l'utilisateur.

L'approche considérée dépasse l'approche de la simple dichotomie pêche/non pêche, et permet d'intégrer l'information présente dans les deux phases de la pêche thonière que sont la route et la recherche active.
Cette cartographie a permis, à partir d'hypothèses sur les trois comportements d'un thonier durant son activité, de produire des indices d'abondance du thon dans la zone couverte (Walker *et al.*, 2015). De même, à partir du modèle qu'ils avaient développé, Joo *et al.* (2013a) ont établi un indice d'abondance de l'anchois du Pérou. Cet indice a ainsi pu être comparé aux campagnes scientifiques d'estimation de l'abondance effectuées dans la zone par techniques acoustiques (Joo, 2013, Chap. 7)

Au final, dans ces approches la trajectoire initialement observée est résumée à une séquence d'activités, qui permet d'obtenir la carte d'activité souhaitée, par une agrégation. La notion de mouvement présente dans les données initiales de trajectoires a complètement disparu pour la construction des cartes.
De manière générale l'agrégation sur une grille nécessite de donner un poids identique à chaque succession de points. Par conséquent, il faut nécessairement un échantillonnage régulier de la trajectoire, quitte à faire une interpolation linéaire de celle ci, comme fait par Russo *et al.*

(2014b). Ainsi, la carte est construite *a posteriori* à partir de résumés des trajectoires, mais n'est pas considérée *a priori* comme une des causes du mouvement.

Une autre vision que celle proposée ci dessus serait de décrire les trajectoires comme conséquence d'un champ jouant le rôle de mécanisme moteur de la trajectoire. En adoptant des formalismes similaires à ceux utilisés en écologie animale (section 1.4.2), on peut espérer dresser une carte qui guide le déplacement. Cette carte peut alors être vue comme une carte de préférence de l'environnement, combinaison de différents facteurs (abondance en ressource, zones réglementaires,...). Un tel modèle s'inspirerait des travaux évoqués plus en écologie animale, tels que ceux de Brillinger (2010). Le but de l'estimation devient alors, à partir des trajectoires observées, de reconstruire ce potentiel attractif de la zone d'étude.

1.6 Cas d'étude

1.6.1 La Manche Est

> *« [L]a diversité et l'abondance des ressources marines vivantes font de la [Manche Est] un secteur de pêche économiquement important pour les flottilles artisanales qui y trouvent de nombreuses espèces de poissons, crustacés, mollusques... menacées par la surexploitation et les modifications de l'environnement. Leurs zones de ponte et de nourricerie, leurs voies de migration peuvent être aussi fortement perturbées par les projets d'extraction de sables et graviers, l'installation d'éoliennes en mer, la pose de câbles sous-marins, la pollution... Il est donc essentiel que la communauté scientifique fournisse aux structures décisionnelles les connaissances nécessaires à une meilleure gestion de ces ressources vivantes et de leur exploitation. »*
> **Andre *et al.* (2009)**

La Manche Est (figure 1.4) est une zone à forts enjeux économiques en Europe. Elle est le lieu de nombreuses activités humaines aux intérêts parfois conflictuels : tourisme et loisirs, ports internationaux et fret, exploitation de ressources (vivantes ou non) (Andre *et al.*, 2009).
Cette zone concentre une forte activité pour la pêche, partagée entre les flottes françaises, anglaises, néerlandaises, belges, allemandes, danoises et irlandaises. La flotte française y étant la plus importante en termes de poissons débarqués (avec 56 % des débarquements totaux venant de cette zone entre 2002 et 2010, (Girardin, 2015)). La façade de la Manche Est comptait 820 navires immatriculés français en 2012 (dont certains n'opèrent cependant qu'en Mer du Nord), dont 236 de plus de 12 mètres, donc sujets à la réglementation VMS (Leblond *et al.*, 2014a).

Les navires de plus de 12 mètres sont majoritairement des navires pêchant à plus de 12 miles des côtes. les données utilisées dans cette thèse sont principalement issues de la flottille des navires qui utilisent un engin traînant ciblant des espèces de fond (Leblond *et al.*, 2014a). Ces engins sont des chaluts à panneaux, des chaluts à perche, et des dragueurs.

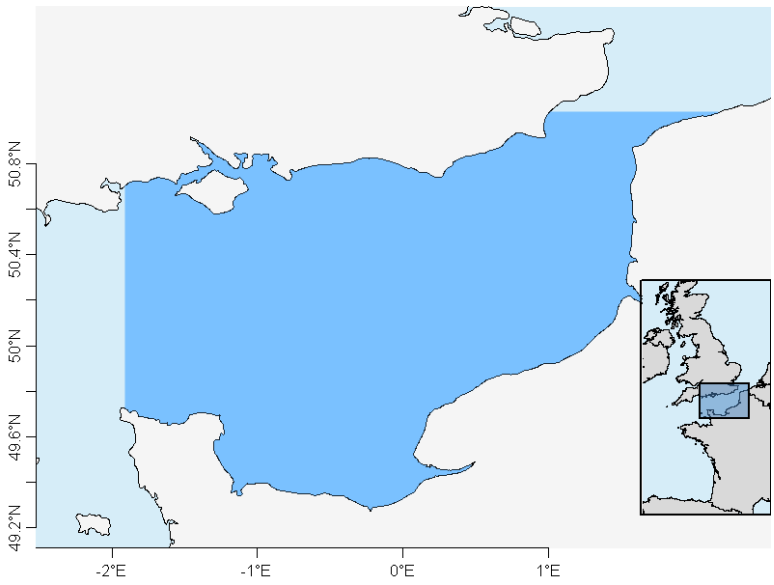


FIGURE *1.4 – La Manche Est (ou carré CIEM VIId)*

Cette flottille est la flottille dominante en Manche Est pour les navires sujets à la réglementation VMS, puisqu'elle représente plus de 80% des navires de cette catégorie.
Les principales espèces ciblées en Manche Est par les navires utilisant ces engins sont la sole, la plie, le bar, la seiche, le merlan ou la coquille saint Jacques.

1.6.2 Un complément aux VMS : Le projet RECOPESCA

Parallèlement au système réglementaire VMS, IFREMER a développé en partenariat avec des pêcheurs volontaires, le projet RECOPESCA (Leblond *et al.*, 2010). Ce projet consiste en l'installation de capteurs sur les navires de pêche, dont des capteurs GPS. Le suivi des navires par le projet RECOPESCA apporte plusieurs avantages par rapport aux données VMS réglementaires :

1. Le pas de temps de l'acquisition GPS est beaucoup plus fin que celui des données VMS, il est de 15 minutes ;

2. L'acquisition est beaucoup plus régulière que celle des données VMS.

Nous disposions du suivi de cinq navires dans la Manche Est, dont les données ont été utilisées pour l'étude menée dans le chapitre 2. Certains engins de pêche utilisés par les navires participant au projet RECOPESCA disposent également de capteurs de mise à l'eau, permettant la connaissance du moment des opérations de pêche. Ces données offrent la possibilité de valider les inférences réalisées pour détecter les moments d'activités de pêche. Ces données ont notamment été utilisées dans une étude des performance de modèles de segmentation pour l'analyse de trajectoires (Bez *et al.*, prep). Certains résultats de cette étude sont montrés à la fin du chapitre 2.

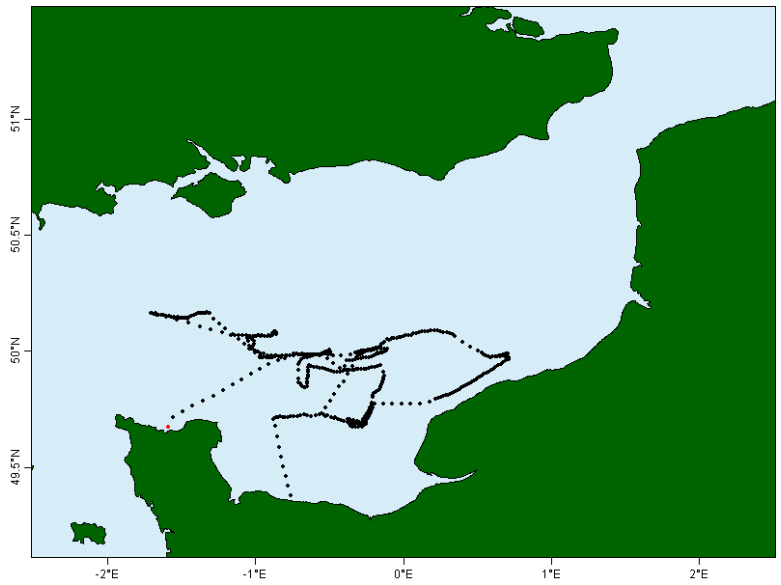


FIGURE *1.5 – Exemple d'observation d'une trajectoire d'un chalutier de fond en Manche Est. Le navire part de Cherbourg (point rouge) pour revenir à Port-en-Bessin. La marée dure 6 jours et donnent lieu à 505 acquisitions GPS. Trajectoire issue du projet RECOPESCA.*

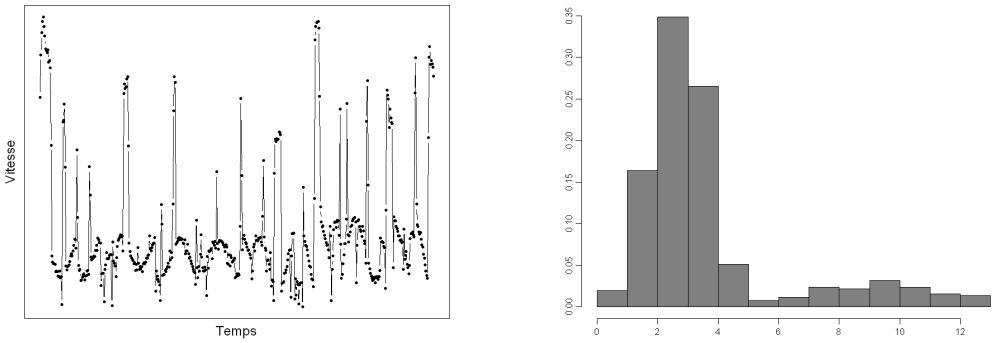


FIGURE *1.6 – Vitesse calculée d'un chalutier de fond à partir des positions de la figure 1.5. À gauche, la série temporelle, à droite, l'histogramme associé. Deux régimes de vitesses semblent être adoptés.*

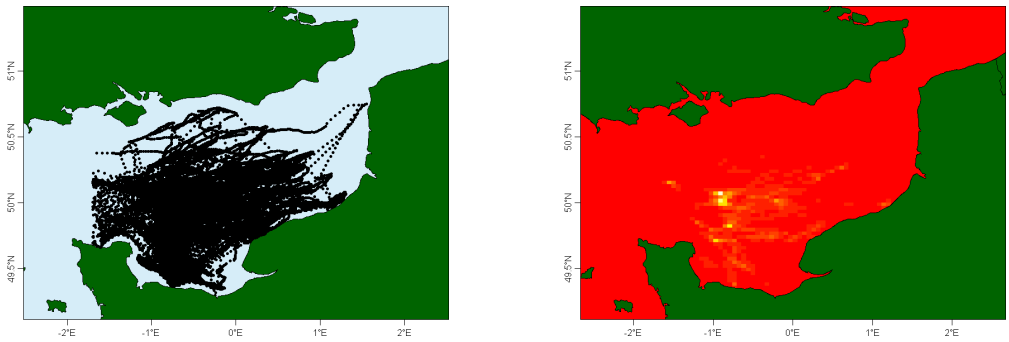
1.6.3 Un premier aperçu des problématiques

Plusieurs comportements au cours d'une trajectoire

La figure 1.5 représente la trajectoire d'un chalutier de fond en Manche Est. Le calcul des vitesses à partir des positions observées montre que le navire semble adopter deux comportements. Le premier semble se caractériser par un régime de vitesse élevé, avec une vitesse supérieure à 7 nœuds, tandis que le second semble être caractérisé par une vitesse plus faible, entre 2 et 4 nœuds. (figure 1.6). La problématique de la détection de ces régimes le long de la trajectoire, à partir d'un modèle de mouvement, sera développée dans le chapitre 2.

Définir un champ spatial à partir des trajectoires

La figure 1.7 représente un groupe d'observations de trajectoires d'un chalutier de fond en Manche Est (figure 1.7a). En projetant ces observations dans une grille de taille prédéfinie, on a obtenu un champ spatial d'utilisation de la Manche par ce navire 1.7b. Le chapitre 3 traitera de différentes méthodes statistiques pour obtenir de telles cartes. De plus, nous y développerons



(a) Observations *(b) Agrégation*

FIGURE *1.7 – Gauche : Exemple d'observation d'un groupe de trajectoires d'un chalutier de fond en Manche Est (trajectoires issues du projet RECOPESCA). Droite : Représentation d'un champ spatial à partir des trajectoires. Ce champ est obtenu en comptant le nombre de positions observées dans une grille de taille prédéfinie. Les zones rouges ne contiennent aucun point, les zones blanches en contiennent beaucoup.*

un modèle de mouvement décrivant les observations de trajectoire comme la conséquence d'un champ spatial de perception de l'environnement par les pêcheurs.

1.7 Objectif de la thèse et démarche adoptée

Le travail présenté dans ce manuscrit propose d'étudier la pertinence d'une modélisation mécaniste des trajectoires de navires de pêche pour l'halieutique. La démarche adoptée entrera dans le cadre décrit dans la section 1.2.3. Le mouvement des navires de pêche est vu comme la conséquence de différents phénomènes que les trajectoires observées doivent permettre d'estimer. Les modèles développés sont ainsi conformes au paradigme voulu par Nathan *et al.* (2008), même s'ils n'en englobent pas tous les compartiments. Le cadre de modélisation adopté sera un cadre de modélisation stochastique, utilisant le formalisme des probabilités. Les modèles développés ici ont pour objectif d'être appliqués à des trajectoires réellement observées en halieutique, mais aussi en écologie. Une grande importance sera donc apportée à l'explicitation des méthodes d'inférence pour ces modèles.

Le chapitre 2 est consacré au développement d'un modèle de Markov caché pour détecter l'activité de pêche à partir de trajectoires de navires. Les modèles HMM y seront décrits, ainsi que les techniques d'inférence associées. On replacera le modèle dans le contexte de la segmentation en écologie du mouvement. Une large place sera accordée à la discussion de la performance du modèle, et à ses limites. Ainsi, on s'appuiera sur les modèles HMM pour discuter de l'intérêt de considérer différentes mesures alternatives de la vitesse (e.g. vitesse par rapport au sol, vitesse par rapport à la masse d'eau) comme proxy pour segmenter les diverses activités (e.g. pêche, route...) au cours d'une marée.

Le chapitre 3 propose un changement complet de formalisme, et décrit un modèle continu en temps et en espace pour décrire le déplacement en écologie. Le modèle s'appuie sur le formalisme

des équations différentielles stochastiques (EDS). Ce formalisme sera présenté succinctement. Les problématiques de l'inférence de ces modèles seront exposées. Nous verrons qu'elles sont proches de celles des modèles HMM. Le modèle sera ensuite décrit, ainsi que son cadre d'inférence, et appliqué à notre jeu de données pour estimer le champ de potentiel latent. Ce champ sera ensuite comparé aux cartes d'abondance de la ressource estimée par une campagne scientifique. Cette étude permettra de tester l'hypothèse selon laquelle l'activité de pêche se concentre sur les zones estimées comme étant à forte abondance en ressources. La pertinence des modèles continus pour l'halieutique et l'écologie sera ensuite discutée.

Enfin, le chapitre 4 offrira une discussion générale sur l'intérêt de la modélisation mécaniste pour l'étude des trajectoires en halieutique. Une large part de ce chapitre sera consacrée à la discussion des principales pistes de recherche ouvertes par ce travail, pour l'halieutique comme pour l'écologie du mouvement.

Chapitre 2

Reconstruire une séquence de comportements à partir de la trajectoire : Une approche par modèle de Markov caché

2.1 Problématique de la détection de comportements

En écologie comportementale, il est classique de s'intéresser à la succession de comportements différents adoptés par un individu durant son cycle de vie. Ces comportements sont le fruit d'un changement dans l'état interne, qui se traduit par undans l'activité de l'individu (Nathan *et al.*, 2008). On suppose donc qu'au cours du temps, un individu adopte une suite de comportements distincts. L'ensemble des comportements est considéré comme un ensemble discret. Par exemple, on dira qu'un animal est en recherche de nourriture, au repos, en migration... Le postulat de base en écologie du mouvement est que deux comportements différents induisent deux manières différentes de se mouvoir (Getz and Saltz, 2008). L'analyse des trajectoires d'un individu doit permettre alors de retracer la séquence comportementale sous-jacente à ce déplacement.

Soit une trajectoire[1] $(\mathcal{X}_t)_{0\leq t\leq T}$, échantillonnée aux temps $t_0 = 0, \ldots, t_n = T$, par les observations $X_0, \ldots, X_n$. Dans tout ce chapitre, on considère que l'observation est faite sans erreur. C'est à dire que $\mathcal{X}_i = X_i$ pour chaque observation (voir section 1.2.1). On considère qu'il existe parallèlement à ce processus, un processus $(\mathcal{S}_t)_{0\leq t\leq T}$ à valeurs dans un ensemble discret $\mathcal{D}$, qui décrit le comportement adopté par l'individu au cours de la trajectoire. Conjointement à la séquence $X_0, \ldots, X_n$, il existe donc une séquence $S_0, \ldots, S_n$ de comportements adoptés par l'individu aux temps $t_0, \ldots, t_n$.

On s'intéresse alors à décrire le couple (X_t, S_t). Deux questions majeures se posent pour décrire ce couple :

1. Quelles caractéristiques de la trajectoire observée sont les plus susceptibles de révéler le

1. Avec les notations de la section 1.2.1

comportement ?

2. Quelles hypothèses fait on sur l'ensemble des comportements possibles, et sur la structure de dépendance dans leur succession au cours du temps ?

2.1.1 Quelle série temporelle pour détecter les comportements ?

Pour détecter les comportements sous-jacents de l'individu, il convient de se poser la question suivante : la série des positions géographiques d'un individu est elle le meilleur proxy de ses comportements ? Autrement dit, pour deux comportements distincts notés $S = 1$ ou $S = 2$, existe-t-il une ou plusieurs variables (résumées des observations), notées $Y_0, \ldots, Y_n$, telles que la différence entre les couples $(Y_t, S_t = 1)$ et $(Y_t, S_t = 2)$, soit plus marquée, et donc aisée à détecter que celle entre les couples $(X_t, S_t = 1)$ et $(X_t, S_t = 2)$?
Un exemple simple est montré sur la figure 2.1, où on a simulé une trajectoire comme étant la succession de 3 marches aléatoires Gaussiennes [2] dont les paramètres de diffusion vont croissant avec le temps. Dans cet exemple, on choisit d'extraire des observations quatre séries temporelles différentes : les positions (Longitude/Latitude), la vitesse scalaire, la direction, et les changements de direction. Ces trois dernières séries sont calculées à partir des positions. La figure illustre que segmenter la trajectoire observée en 3 morceaux homogènes (i.e., 3 marches aléatoires de diffusion constante) semble "de visu", plus facile en ne considérant pas directement les observations, mais un *résumé* de celles-ci, la vitesse moyenne entre chaque position.
Cet exemple illustre que pour détecter les comportements, le choix de variables résumées peut être primordial. Dans la suite, on désigne par $(Y_t)_{t=0,\ldots,n}$ la série utilisée pour détecter les comportements constitué de la trajectoire, et/ou d'une ou plusieurs de ses dérivées. Plusieurs séries dérivées de la trajectoire ont été utilisées dans la littérature. La (ou les) série(s) choisies doi(ven)t refléter les comportements d'intérêt pour l'écologue, et satisfaire les hypothèses statistiques nécessaires à la technique de segmentation utilisée. À partir de données de trajectoires, les séries dérivées les plus classiquement utilisées sont :

La série des normes des vitesses $(V_i)_{i=1,\ldots,n}$

$$V_i = \frac{\| X_i - X_{i-1} \|}{t_i - t_{i-1}} \tag{2.1}$$

Cette série est à valeur dans $\mathbb{R}_+$. Elle est souvent utile pour dissocier des comportements agissant sur l'amplitude des déplacements (migration ou fuite pour de longs déplacements, sédentarisation ou repos, pour des déplacements courts). Un exemple est montré sur la figure 2.1. Ce résumé fait l'hypothèse que l'individu se déplace en ligne droite entre les deux points observés. Cette approximation sous-estime donc la vitesse de l'individu, d'autant plus que le

2. Une marche aléatoire Gaussienne de paramètre de diffusion σ^2 est un processus $(X_t)_{t\geq 0}$ tel que

$$X_0 = X_0, \quad X_{t+\Delta} = X_t + \varepsilon_\Delta, \quad \varepsilon_\Delta \sim \mathcal{N}(0, \Delta\sigma^2)$$

temps d'acquisition entre les positions est long (voir par exemple Hintzen *et al.*, 2010, pour l'impact sur les vitesses des navires de pêche).

La série des directions $(\phi_i)_{i=1,\dots,n}$
En posant $X_i = (x_i, y_i)^T$

$$\phi_i = \arg z_i, \text{ où } z_i := (x_i - x_{i-1}) + j(y_i - y_{i-1}) \in \mathbb{C}$$

Cette série est à valeur dans $]-\pi;\pi]$. Elle est utile quand les comportements agissent sur l'orientation du mouvement. Cela permet d'intégrer l'attraction vers une zone, une proie ou un congénère, ou la répulsion d'une zone ou d'un prédateur. Un exemple est montré en annexe sur la figure E.1. Encore une fois, on fait l'hypothèse que l'individu se déplace en ligne droite entre les deux points observés. Cette série est définie sur le cercle, son caractère circulaire requiert donc donc une modélisation particulière.

La série des changements de direction $(\psi_{t_i})_{i=2,\dots,n}$

$$\begin{aligned} \psi_i &= \phi_i - \phi_{i-1} && \text{si } |\phi_i - \psi_i\phi_{i-1}| \leq \pi \\ \psi_i &= (\pi - (\phi_i - \phi_{i-1})) \times (-1)^{\mathbf{1}\{\phi_i - \phi_{i-1} < -\pi\}} && \text{si } |\phi_i - \phi_{i-1}| > \pi. \end{aligned}$$

Où $\mathbf{1}\{E\}$ vaut 1 si l'évènement E est vrai, 0 sinon. Cette série prend valeur entre $-\pi$ et π (on considère qu'un changement de direction ne dépasse jamais un demi tour). Elle est utile si les comportements sont associés à des portions de mouvement plus ou moins rectilignes (mais sans orientation fixe, comme dans le cas précédent). Par exemple, la route d'un prédateur vers une zone de chasse va être plus rectiligne que la recherche à l'intérieur de cette zone. Un exemple est montré en annexe sur la figure E.2. Cette série est l'approximation discrète d'une dérivée seconde. L'erreur faite dans ce calcul peut par conséquent être plus grande que pour les deux séries présentées plus haut.

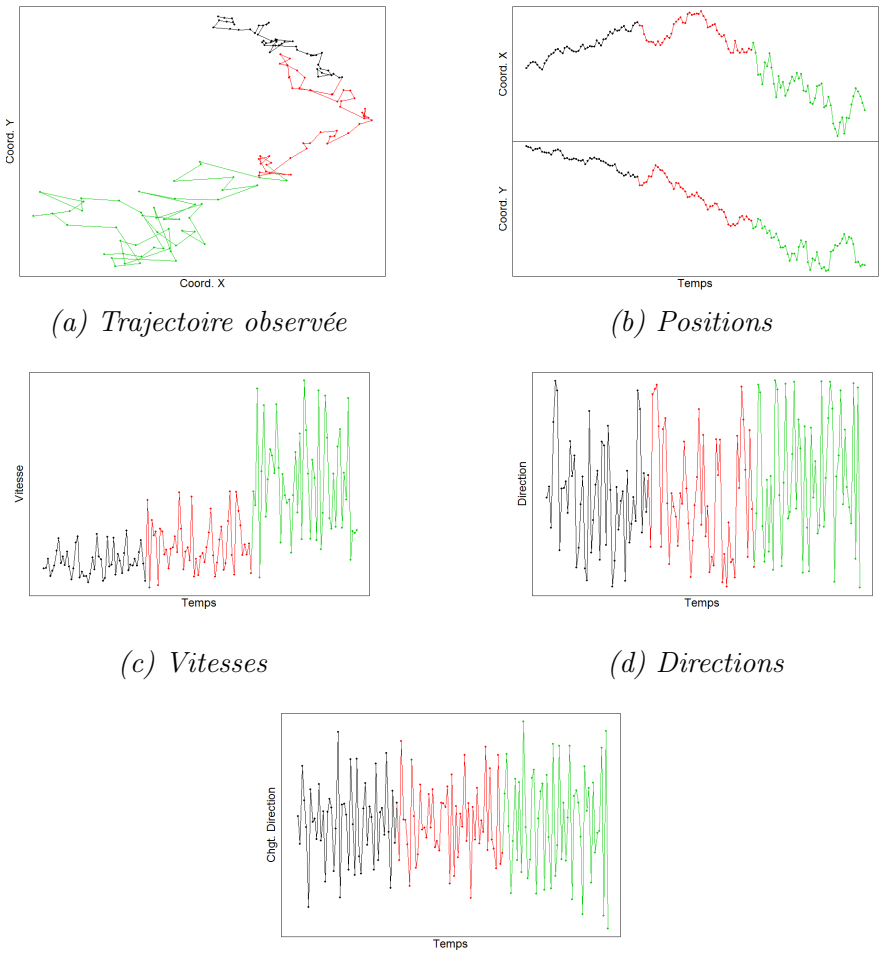
2.1.2 Quelle structure pour la séquence des comportements ?

De manière générale, on considère que pour un comportement s, le mouvement d'un individu est régi par un ensemble de paramètres θ_s. On dit donc que deux comportements s_1 et s_2 sont différents si $\theta_{s_1} \neq \theta_{s_2}$. Une fois choisie une série $(Y_t)_{t=0,\dots,n}$ [3], une autre question essentielle réside dans les hypothèses concernant

1. La structure de l'espace des comportements ;
2. La structure de dépendance dans la séquence des comportements.

On peut dissocier ici deux types d'approches, qui se distinguent par leur définition des deux points ci dessus.

3. Cette série peut être l'observation X_t elle même



(a) Trajectoire observée *(b) Positions*

(c) Vitesses *(d) Directions*

(e) Changements de direction

FIGURE *2.1 – Cinq différents points de vue d'un mélange de marches aléatoires Gaussiennes. Le paramètre de variance va du faible (noir) au fort (vert). Les observations sont représentées par (a) la trajectoire et (b) la série bivariée des positions. Trois résumés de la trajectoire sont représentés : (c) la vitesse ,(d) les directions et (e) les changements de direction. Pour retrouver les comportements, il semble plus facile ici de s'intéresser à la série des vitesses, et non à la trajectoire en elle même.*

Détection de ruptures

Dans l'approche par détection de ruptures, la démarche est la suivante. On suppose que la série $(S_t)_{t=0,\dots,n}$ est affectée par $K-1$ $(K \leq n)$ ruptures. Ainsi, cette séquence prend K valeurs (ou comportements) distinctes. Comme dit plus haut, ces K comportements distincts se traduisent par K différents paramètres régissant le mouvement. Étant donnés $(Y_i)_{i=0,\dots,n}$, on pose un modèle paramétrique tel que

$$Y_t \sim f(\cdot, \theta).$$

On suppose que le paramètre θ est affecté par $K-1$ changements advenus aux temps $t_1, \dots, t_{K-1}$ (en posant $t_0 = 0,\ t_K = t_n = T$), tels que

$$\forall t \in I_k, \quad Y_t \sim f(\cdot, \theta_k) \tag{2.2}$$

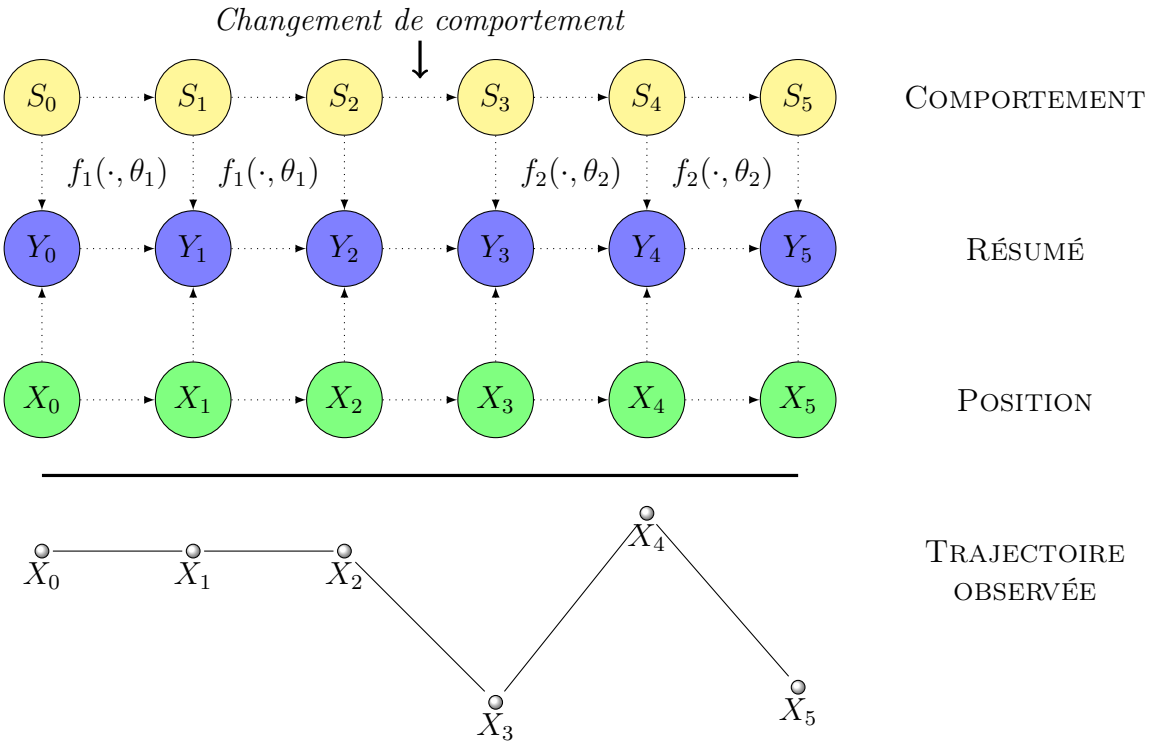


Figure *2.2 – Représentation hiérarchique de la relation comportement/trajectoire. Le processus des comportements influe sur les positions (de par son influence sur le (ou les) résumé(s) de celles-ci). Un changement de comportement induit un mouvement différent (ici d'abord linéaire, puis erratique).*

où $I_k =]t_{k-1}; t_k]$ est l'intervalle de temps où $\theta = \theta_k$ [4].

Dans la pratique, K est connu ou inconnu. Dans le cas où le nombre de ruptures est connu, les instants de rupture sont détectés en maximisant la vraisemblance du modèle, en utilisant des algorithmes de programmation dynamique. Quand K est inconnu, on applique cette méthode pour différentes valeurs. K est ensuite sélectionné selon un critère de sélection de modèles (Lebarbier, 2005, par exemple).

La majorité des études considèrent que, pour un segment temporel I_k, les $(Y_t)_{t \in I_k}$ sont indépendants. Les méthodes utilisées ont récemment été étendues aux processus autoregressifs (Chakar *et al.*, 2014). L'utilisation de ces méthodes pour détecter différents comportements a été proposée pour des mammifères marins (Gurarie *et al.*, 2009), et des oiseaux (Madon and Hingrat, 2014).

Dans cette approche, un même comportement n'est pas adopté sur deux segments temporels distincts. En ce sens, il existe autant de comportements que de segments homogènes, et la structure de transition est irréversible.

Des extensions de ces méthodes, palliant ce problème, ont été proposées. L'ensemble final $\{\theta_k, k = 1, \ldots, K\}$ est réduit à un ensemble $\{\theta_j, j = 1, \ldots, J, J \leq K\}$, par méthode de classification (supervisée, ou non, Picard *et al.*, 2007). Dans cette méthode, aucune structure n'est cependant proposée pour la séquence des comportements.

4. Notations reprises de la thèse de Franck Picard, (Picard, 2005)

Les modèles à espace d'état

Une alternative à cette approche est l'utilisation de modèles à espace d'états (State Space Models, SSM). On considère que la série $S_0, \ldots, S_n$ est à valeurs dans l'espace fini (de cardinal connu) des comportements $\mathcal{D}$, appelé espace d'états. Le cardinal de l'ensemble $\mathcal{D}$ est alors connu, et ses éléments peuvent être définis dans l'idée qu'ils correspondent à un ensemble interprétable (migration, recherche, activité de pêche). On considère alors que la distribution de probabilité de Y_m dépend du comportement S_m que l'individu avait adopté au temps t_m. Soit un comportement j, élément de l'ensemble des comportements $\{1, \ldots, J\}$, et un modèle paramétré par θ. On suppose alors, qu'au temps t_m

$$Y_m | \{S_m = j\} \sim f_j(\cdot, \theta) \tag{2.3}$$

On a ainsi un schéma hiérarchique dans lequel un état caché impacte un processus observé (les $(Y_t)_{t=0,\ldots,n}$), comme montré sur la figure 2.2.
Malgré une écriture différente, l'équation (2.3) est très proche de l'équation (2.2). Une différence est que le cardinal de l'espace d'états $\mathcal{D}$ est nécessairement connu *a priori*. Une autre différence est la structure des transitions dans la séquence $S_0, \ldots, S_n$ des états. On ne suppose pas ici qu'un état ne puisse être revisité. L'individu peut donc adopter un comportement, et y revenir plus tard.
Un atout majeur pour la modélisation est la possibilité de spécifier un modèle de transition entre états, en posant

$$\mathbb{P}(S_{m+1} = j | S_0^m = s_0^m, \theta) = p_{s_0^m, j}(t_m) \tag{2.4}$$

où S_0^m est la séquence $S_0, \ldots, S_m$ des comportements aux temps $t_0, \ldots, t_m$ et s_0^m est un élément de l'ensemble $\{1, \ldots, J\}^{m+1}$.
Une telle modélisation implique une structure de dépendance dans la séquence des états. L'intégration de telles structures doit permettre

1. de donner une cohérence à la séquence des états ;

2. d'expliciter les mécanismes de transition dans les comportements des individus, et ainsi aider à mieux décrire le mouvement observé.

Un exemple simple d'une telle structure est celle d'indépendance des états, où

$$\mathbb{P}(S_{m+1} = j | S_m = i, \theta) = \mathbb{P}(S_{m+1} = j | \theta) = p_j(t_m). \tag{2.5}$$

Le modèle sur Y est donc dans ce cas un modèle de mélange. Si, dans l'équation 2.5, la fonction p_j est ne dépend pas du temps t, on dit que le modèle de mélange est homogène en temps.
Un autre exemple classique de structure de transition est une chaîne de Markov. On suppose alors que la probabilité de se trouver dans un état au temps t_{m+1} ne dépend que de l'état au

temps t_m et du temps t_m :

$$\mathbb{P}(S_{m+1} = j | S_m = i, S_0^{m-1} = s_0^{m-1}, \theta) = \mathbb{P}(S_{m+1} = j | S_m = i, \theta) = p_{ij}(t_m), \forall s_0^{m-1} \in \{1, \ldots, J\}^m \tag{2.6}$$

Si la probabilité p_{ij} ne dépend pas du temps, on dit que la chaîne de Markov est homogène en temps. Quand la chaîne des comportements suit un chaîne de Markov, le modèle hiérarchique associé est alors appelé Modèle de Markov Caché (ou Hidden Markov Model, HMM). De nombreuses applications de ces modèles ont été proposées en écologie (Morales *et al.*, 2004; Franke *et al.*, 2006; Holzmann *et al.*, 2006; Jonsen *et al.*, 2007; Patterson *et al.*, 2009, et d'autres encore...) et en halieutique (Vermard *et al.*, 2010; Walker and Bez, 2010; Peel and Good, 2011; Charles *et al.*, 2014).

2.1.3 Quelle méthode pour estimer les comportements ?

Un problème de données manquantes

Soient $\mathbf{Y}$ et $\mathbf{S}$ les vecteurs $(Y_0, \ldots, Y_n)$ et $(S_0, \ldots, S_n)$. Un modèle à espace d'états est donc un modèle paramétré par θ, où on a défini la loi de $\mathbf{Y}$ sachant $\mathbf{S}$ (par l'équation (2.3)) ainsi qu'une structure de transition dans l'espace d'états (du type (2.5), ou (2.6)). Dans un tel modèle, la séquence des états comportementaux adoptés durant la trajectoire est non observée. Il s'agit en ce sens d'une donnée manquante.
À partir d'une telle modélisation, il existe un double objectif d'estimation.

1. Déterminer le maximum de vraisemblance des paramètres du modèle

$$\hat{\theta} = \text{argmax}_\theta \, p(\mathbf{Y}|\theta)$$

2. Déterminer la séquence de comportements la plus probable correspondant à ce maximum de vraisemblance

$$\hat{\mathbf{S}} = \text{argmax}_{\mathbf{S}} \, p(\mathbf{S}|\mathbf{Y}, \hat{\theta})$$

La structure hiérarchique des modèles à espace d'états permet ici d'écrire la vraisemblance :

$$\begin{aligned} L(\theta) := p(\mathbf{Y}|\theta) &= \sum_{\mathbf{S}} p(\mathbf{Y}, \mathbf{S}|\theta) \\ &= \sum_{\mathbf{S}} p(\mathbf{Y}|\mathbf{S}, \theta) p(\mathbf{S}|\theta) \end{aligned} \tag{2.7}$$

Pour obtenir le maximum de vraisemblance, il faut maximiser l'équation (2.7). L'atteinte de cet objectif ne peut pas toujours se faire de manière directe [5], en fonction de la nature du modèle. Il est alors nécessaire d'utiliser des algorithmes spécifiques, permettant de maximiser la vraisemblance en explorant l'espace de tous les états cachés du système. Un algorithme classique dans le cadre présenté ici est l'algorithme EM.

5. Pour des raisons théoriques ou pratiques

L'algorithme EM

De manière générale, l'algorithme Expectation Maximization (EM) est un algorithme permettant de maximiser la vraisemblance d'un modèle dont les sorties ne sont que partiellement observées (Dempster *et al.*, 1977).
Pour un modèle paramétré par θ, étant données des observations $\mathbf{Y}$, on veut maximiser la fonction de vraisemblance $L(\theta)$ telle que définie dans l'équation (2.7). Cependant, lorsque cette vraisemblance est non analytique, ou impossible à calculer en pratique [6], il est possible de maximiser L indirectement, grâce à un algorithme itératif.
Dans les modèles à espaces d'états tels que définis en section 2.1.2, on appelle données manquantes [7], le vecteur $\mathbf{S}$. On appelle alors "données complètes" le couple $(\mathbf{Y}, \mathbf{S})$. Si on note $l(\theta) = \log L(\theta)$, on peut écrire

$$l(\theta) = \log\left(p(\mathbf{Y}, \mathbf{S}|\theta)\right) - \log\left(p(\mathbf{S}|\mathbf{Y}, \theta)\right) \tag{2.8}$$

Soit un ensemble de paramètres θ_0 quelconque, alors, en intégrant (2.8) par rapport à la densité $p(\mathbf{S}|\mathbf{Y}, \theta_0)$, on obtient

$$l(\theta) = \underbrace{\mathbb{E}_{\mathbf{S}|\mathbf{Y},\theta_0}\left(l(\mathbf{Y}, \mathbf{S}|\theta)\right)}_{:=Q(\theta,\theta_0)} - \mathbb{E}_{\mathbf{S}|\mathbf{Y},\theta_0}\left(l(\mathbf{S}|\mathbf{Y}, \theta)\right) \tag{2.9}$$

La fonction $Q(\theta, \theta_0)$ est donc l'espérance de la log vraisemblance complète de θ, sous la loi des données manquantes conditionnellement aux observations et à θ_0. On peut montrer que

$$Q(\theta, \theta_0) \geq Q(\theta_0, \theta_0) \Rightarrow L(\theta) \geq L(\theta_0) \tag{2.10}$$

La relation (2.10) met sur la piste d'un algorithme itératif pour maximiser $l(\theta)$. En effet, pour augmenter la vraisemblance, il suffit d'augmenter, en θ, la fonction $Q(\theta, \theta_0)$. L'algorithme EM (algorithme 1) est un algorithme itératif basé sur ce principe. La suite des $(\theta_k)_{k\in\mathbb{N}}$ converge ainsi

Algorithme 1 Algorithme EM

1: Choisir θ_0
2: **Pour** k allant de 1 à K
3: ÉTAPE E Calculer $Q(\theta, \theta_0)$ comme dans l'équation (2.9)
4: ÉTAPE M Calculer $\theta_k = \text{argmax}_\theta Q(\theta, \theta_{k-1})$
5: **Fin**

vers un maximum [8] de la vraisemblance (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997).

L'étape E peut s'avérer complexe et est grandement dépendante du modèle. Dans le cadre des HMM, elle peut s'effectuer à l'aide d'un algorithme de programmation dynamique. La conjonction de l'algorithme EM avec cet algorithme est connue sous le nom d'algorithme de Baum

6. pour des raisons de temps de calcul
7. Au sens : non observées
8. Local, évidemment, il faut qu'il reste un challenge !

Welch (Baum *et al.*, 1970). Un exemple d'implémentation de l'algorithme de Baum Welch, pour le modèle développé en section 2.2 est entièrement décrit en annexe A.2.
Dans les cas où la fonction Q est impossible à calculer analytiquement, on peut approcher l'espérance par méthode de Monte Carlo. Sous θ_0, on simule les données manquantes conditionnellement aux observations. La fonction Q est alors approchée par la moyenne empirique de la vraisemblance complète. Cette variante de l'algorithme EM est appelée Monte Carlo (MC)EM (Wei and Tanner, 1990). Dans ce cas, on simule un nouveau jeu de données manquantes à chaque itération. Une alternative à l'algorithme MCEM a été proposée, appelée approche SAEM (Delyon *et al.*, 1999). L'étape E est toujours approchée par simulation, cependant, elle est intégrée sur toutes les itérations, pour limiter le nombre de données manquantes à simuler. Cette approche est particulièrement efficace quand le temps de calcul de l'algorithme est essentiellement dû à la simulation, et non à la maximisation.
L'étape M, dans certains modèles, peut être directe. Cependant, lorsque la fonction $Q(\theta, \theta_0)$ a une forme complexe (non convexe, multimodale), obtenir le maximum global peut représenter un problème sophistiqué. Des algorithmes numériques peuvent alors être utilisés.

Pour le cadre proposé sur la figure 2.2, l'algorithme EM permet donc d'estimer les paramètres du mouvement θ. On peut ensuite estimer la séquence des comportements conditionnellement à l'estimateur du maximum de vraisemblance $\hat{\theta}$. Dans le cadre des modèles HMM, la séquence des comportements la plus probable [9] est retracée à l'aide d'un algorithme de programmation dynamique, l'algorithme de Viterbi (Rabiner, 1989). Le détail de cet algorithme est donné dans l'annexe A.2.

9. étant donnés les observations et le maximum de vraisemblance

2.2 Un modèle de Markov caché autoregressif pour décrire l'activité des navires de pêche

Cette section traite de l'utilisation d'un modèle de Markov caché pour décrire l'activité des navires de pêche à partir des observations de leurs trajectoires. Il a donné lieu à un article dans la revue Environmetrics, écrit en collaboration avec Stéphanie Mahévas, Etienne Rivot, Mathieu Woillez, Jérôme Guitton, Youen Vermard et Marie Pierre Etienne.
L'article est ici reproduit tel qu'il a été publié, les méthodes et les résultats y sont ainsi présentés. L'auteur présente ses excuses au lecteur pour le passage à l'anglais [10].

Research Article **Environmetrics**

Received: 5 September 2013, Revised: 17 September 2014, Accepted: 9 October 2014, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2319

An autoregressive model to describe fishing vessel movement and activity

P. Gloaguen[a*], S. Mahévas[a], E. Rivot[b], M. Woillez[c,b], J. Guitton[b], Y. Vermard[d] and M. P. Etienne[e]

2.2.1 Introduction

Understanding the dynamics of fishing vessels is essential to characterize spatial distribution of fishing effort on a fine spatial scale, thus to estimate the impact of fishing pressure on the marine ecosystem (Poos and Rijnsdorp, 2007; Mills *et al.*, 2007), or to understand fishermen's reactions to management measures (Vermard *et al.*, 2008) and to improve the assessment of the impact of management plans (Lehuta *et al.*, 2013). Assessing the spatiotemporal distribution of vessels targeting fish populations also improve the understanding of the dynamics of fish resources (Bertrand *et al.*, 2004 ; Poos and Rijnsdorp, 2007).

Modelling the dynamics of fishing vessels is classically approached by statistical analyses of landing declarations. This has a low spatial resolution (ICES statistical rectangle) (Hutton *et al.*, 2004 ; Pelletier and Ferraris, 2000). Recently the mandatory Vessel Monitoring System (VMS), for legal controls and safety (Kourti *et al.*, 2005), has led to massive acquisition of fishing vessels' movement data which offer new means of studying fishermen spatio-temporal dynamics. Data consist in geographical positions recorded at a more or less regular time step (less than two hours for mandatory VMS data). The French Institute for the Exploitation of the Sea (IFREMER) has recently developed the RECOPESCA project with volunteer fishermen, whose vessels positions are recorded at a 15 minutes time step.

10. A fortiori, mon anglais

Mechanistic mathematical models have long been used in ecological sciences to analyse movements and activity of different tracked animals (Bovet and Benhamou, 1988; Flemming *et al.*, 2006).

A key issue in behavioral ecology is the identification of the sequence of hidden (non observed) behaviors from the analysis of the trajectory, such as foraging, research, migration. Similar questions are investigated in fisheries science, where the identification of different behaviors adopted by fishing vessels during a fishing trip (route towards fishing zone, fishing activity...) is of interest to understand what drives fishing activities and fishing effort dynamics.

Hierarchical models are well suited to analyse trajectories and hidden sequence of behaviors from discrete positions records. They first describe a non observed time-behavioral process based on different behavioral states adopted by the individual and rules for switching from one to the other. The path is then modelled conditionally to the behavioral state. These models are commonly called State Space Models (SSM) and (when the sequence of hidden state satisfies the Markov property) Hidden Markov Models (HMM) (Patterson *et al.*, 2009; Langrock *et al.*, 2012; Jonsen *et al.*, 2013b). In animal ecology, these models were already used to describe the path of different animals such as elks (Morales *et al.*, 2004) or seals (Jonsen *et al.*, 2005). The application of those tools to infer fishing vessels activity has received comparatively little attention (Vermard *et al.*, 2010, Walker and Bez, 2010, or Peel and Good, 2011).

The model developed in the present paper present two substantial contributions to the models published so far to analyse fishing vessel trajectories.

First, keeping the Markovian structure for the hidden behavior of the vessel, we propose to describe the vessel's path via the modeling of the velocity process. The latter is modeled with a bivariate Gaussian first order autoregressive process (AR) which parameters are conditioned by the behavioral state. Models published so far have used scalar speed and turning angles to describe the vessel's path conditionally upon the behavioral state. Those two variables are usually modelled without autocorrelation between time steps. For instance, Bayesian model developed by (Vermard *et al.*, 2010) define scalar speed and turning angles for each time step as drawn a priori in independent Gaussian and Wrapped Cauchy distributions with parameters specific to the behavioral state. The AR bivariate velocity process allows the use of a unique Gaussian structure instead of two separated distributions for speed and turning angles (Gurarie *et al.*, 2009), and therefore makes easier the modelling of autocorrelation to capture the inertia in the velocity process. This process has already been used in animal ecology to describe animal's velocity, but, to the best of our knowledge, this work represents its first use in fisheries science to describe vessel's path. It also offers computational facilities for the inferences.

Second, the model is fitted to original GPS records data issued from the RECOPESCA project

(Leblond *et al.*, 2010). The RECOPESCA project is implemented by the French institute for the exploitation of the sea (IFREMER) to improve the assessment of the spatial distribution of catches and fishing. Although they concern a rather restricted number of fishing vessels, RECOPESCA data offer several advantages by comparison with mandatory VMS data. First, these data are recorded with a shorter time step than VMS data (a position every 15 minutes instead of 1 hour). Second, they are recorded with a highly regular time step (15 min +/- 1 min). The finer time scale allows for a more accurate reconstruction of fishing vessel trajectories than VMS data. Through an extensive simulation approach, Vermard *et al.* (2010) has already shown that shorter acquisition time step would provide better inferences on the sequence of fishing vessel behavior. Bias induced by interpolating the trajectory with a straight line between two records would be lower than with an hour time step between two points (Skaar *et al.*, 2011). Furthermore, the regularity of these GPS records is essential to formulate the AR process hypothesis.

By contrast with the fisheries-related papers previously published that fostered a Bayesian approach (Vermard *et al.*, 2010, Walker and Bez, 2010), a maximum likelihood estimation approach is adopted. The Baum Welch (BW) algorithm, the expectation-maximization algorithm for Hidden Markov Models, is used to estimate parameters. The BW algorithm is then coupled to the Viterbi algorithm to estimate the hidden behavior sequence. In models previously published (Vermard *et al.*, 2010, Walker and Bez, 2010), the sequence of hidden states has been inferred by using the posterior mode of marginal posterior distribution of each hidden state. By contrast, the Viterbi algorithm achieves estimation of the most credible sequence of behavioral states, that better accounts for the persistence in the sequence stemming from the Markovian property (Rabiner, 1989).

The article is structured in the following way. In section 2, we first detail the RECOPESCA data set. The methodological framework for the HMM coupled with an AR process is then presented. Estimation methods using EM and Viterbi algorithms are later detailed. Our simulation approach used to assess the performance of the estimation method and the sensitivity to alternative data configuration is presented at the end of section 2. Results over simulations and RECOPESCA data are presented in section 3. The ending section proposes a discussion on the adequacy of this modeling approach and some recommendations for future modeling of vessels dynamics.

2.2.2 Material and methods

RECOPESCA data

Four trajectories associated with four different fishing vessels operating in the Channel with different fishing gears are considered to illustrate our modelling approach (see Table 2.1). These four trajectories were extracted from the RECOPESCA data base. For each trajectory, GPS

Trip	Duration	Vessel's length	Gear
A	22h	12m	Dredges
B	14h	12m	Otter Trawl
C	13h	13m	Trammel nets
D	107h	22m	Otter Trawl

TABLE *2.1 – Technical details of the four studied trajectories*

positions in port and at sea were available. As the analysis only focus on fishing vessel movement during fishing trips, we first removed positions in port based on logbooks (landings declarations). The positions were recorded at a regular time step (15 minutes plus or minus 1). Selected trips last more than 12 hours, ensuring enough observed positions for parameters identification. These four vessels belong to the demersal fishery for which the research of fish aggregations observed in pelagic fisheries (such as thuna fisheries, Walker and Bez, 2010) does not exist. Hence only two behaviors are assumed along their path, 'steaming' for cruising and 'fishing' when they operate their gear.

Describing the path with the velocity process

The observed vessel's path, $X_0, \dots X_n$ is considered via a decomposition of the associated velocity process on its two dimensions V^p and V^r.
V^p is called the "persistence" speed, and corresponds to the tendency to persist in the previous direction. V^r is called the "rotational" speed, and corresponds to the tendency to turn. These two quantities are derived as follows :

$$V_t^p = V_t \cos(\psi_t) \tag{2.11}$$

$$V_t^r = V_t \sin(\psi_t) \tag{2.12}$$

where V_t is the average speed derived from positions X_{t-1} and X_t, and ψ_t is the turning angle derived from X_{t-2}, X_{t-1} and X_t, with $\psi_1 = 0$. Variables V^p and V^r model jointly scalar speed and turning angles instead one different distribution for each of them (like in (Vermard *et al.*, 2010) and Walker and Bez, 2010). The bivariate velocity can be modelled using a unique Gaussian structure (Gurarie *et al.*, 2009), that is presented on the next section.
It is worth noting here that the velocity defined by Equations (2.11) and (2.12) is equivalent to the (linearly interpolated) observed trajectory (see the appendix A.1 for the explicit relation).

An AR process ruled by a HMM

The vessel's behavior is modeled by a hidden stochastic discrete time process. This is denoted by $S_0^t := S_0, ..., S_n$, where S_t is the state of the vessel at time t, and takes values in the set of behavioral states noted $\mathcal{S} = \{1, 2\}$, 1 standing for steaming, 2 for fishing.
This process is assumed to be a homogeneous Markov chain of first order with a transition

matrix $\Pi = (\Pi_{ik})_{i,k\in\mathcal{S}}$, i.e :

$$\Pi_{ik} = \mathbb{P}(S_t = k|S_{t-1} = i) = \mathbb{P}(S_t = k|S_0^{t-2}, \quad S^{t-1} = i)$$

The initial distribution is assumed to be known and set to $\mathbb{P}(S_0 = 1) = 1$ (the vessel is steaming when leaving the harbor).

Conditionally to this hidden Markov chain, the velocity process is modeled by a mixture of two dimensional AR processes (with respect to its decomposition in equations (2.11) and (2.12)) and can be summarized as follows :

$$V_{t+1}^p|(S_{t+1} = i) = \ \eta_{p,i} + \mu_{p,i}V_t^p + \sigma_{p,i}\epsilon_{p,t} \tag{2.13}$$

$$V_{t+1}^r|(S_{t+1} = i) = \ \eta_{r,i} + \mu_{r,i}V_t^r + \sigma_{r,i}\epsilon_{r,t} \tag{2.14}$$

$$V_1^p = \ V_1, \quad V_1^r = \ 0, \quad \epsilon_{\cdot,t} \sim \ \mathcal{N}(0,1)$$

where, for each component (V^p or V^r) and state (1 or 2) :

- η is a level parameter.
- μ is an autocorrelation parameter. Its existence is justified considering data from the four different trips (see Figure 2.3 for autocorrelation plots of trips A-D). It's important to note that it is well defined because of the time step regularity.
- σ^2 is a shape parameter, it is the variance of the innovation process.

As in Gurarie *et al.* (2009), processes (2.13) and (2.14) are assumed to be independent. Even if this assumption seems unrealistic, data reveal a weak empirical correlation between those two variables.

AR processes as in (2.13) and (2.14) have a stationary distribution if $|\mu| < 1$ (Shumway and Stoffer, 2000). In this case the expectation and the variance of the process V satisfy asymptotically :

$$\mathbb{E}(V) = \frac{\eta}{1-\mu} \tag{2.15}$$

$$\mathbb{V}(V) = \frac{\sigma^2}{1-\mu^2} \tag{2.16}$$

These asymptotic mean and variance are useful in order to interpret characteristics of the velocity process. If the vessel stays long enough in a given behavior, the expectation for V^p and V^r could be derived from equations (2.15) and (2.16).

Inference

The inference procedure consists in the estimation of both parameters and a reconstruction of the sequence of hidden states from observed positions. It requires two steps 1) performing parameter estimation using the Baum Welch algorithm, 2) estimating the most likely sequence

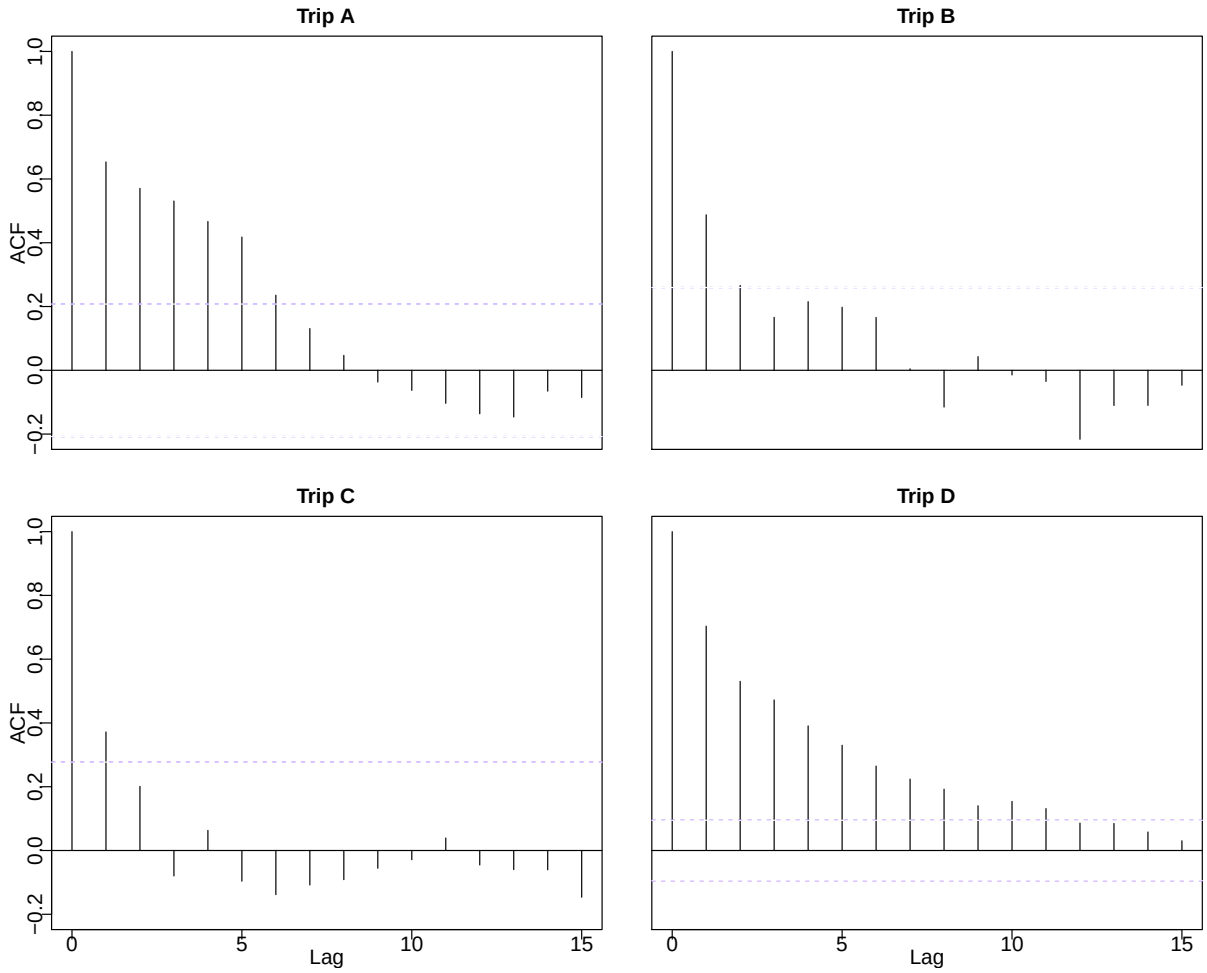


FIGURE *2.3 – Autocorrelation function plots for trips A to D.*

of states using the Viterbi algorithm.
Considering 2 hidden states, the set of parameters to be estimated for this model is

$$\Theta = \{\Pi_{i\cdot},\ \eta_{p,i},\ \eta_{r,i},\ \mu_{p,i},\ \mu_{r,i}\ \sigma_{p,i},\ \sigma_{r,i}\ \}_{i\in\mathcal{S}}$$

Therefore, 14 parameters are estimated (2 for the transition matrix, and $3 \times 2 \times 2$ for AR processes on the velocity). Computing the likelihood function is not possible within a reasonable time as it requires the integration over all possible hidden sequences ($2^{\#\text{Time steps}}$ possible paths). A classical approach is to find Maximum Likelihood Estimators (MLE) $\hat{\Theta}$ via the Baum Welch algorithm, which is the Expectation Maximization (EM) algorithm derived for Hidden Markov Models (Rabiner, 1989 ; McLachlan and Krishnan, 1997).
Considering the model described above, both the Expectation (E) step and the Maximization (M) step can be computed analytically (details and proof are given in the appendix A.2, and **R** codes (R Core Team, 2014) to perform the inference are available on demand).
The EM algorithm is an iterative algorithm to approach the MLE. The convergence criterion is reached when the log likelihood increase is less than 0.01 between two successive iterations. A known problem of the EM algorithm is that, given a starting point, one can converge towards a local maximum of the likelihood. To ensure a global maximum is found, the algorithm was performed from 100 different starting points, keeping the result with the largest likelihood as $\hat{\Theta}$.
Once the MLE step is performed, the Viterbi algorithm is used to derive the most probable sequence of states, accounting for Markovian properties of the whole hidden sequence.

A parametric bootstrap procedure is used to assess the variance of $\hat{\Theta}$. The MLE is used to simulate M new trajectories as bootstrap samples on which MLE $(\hat{\Theta}_m)_{1\leq m\leq M}$ are re-estimated, given these M re-estimations, empirical 95 % confidence intervals are obtained for each parameter (getting central 95% values, McLachlan and Krishnan, 1997; Efron and Tibshirani, 1993). To estimate the uncertainty over the state sequence estimation, the Viterbi algorithm is computed on the derived velocity process using each MLE $\hat{\theta}_m$. Formally, the Viterbi algorithm computes the most probable joint sequence of hidden states and observations :

$$\hat{S}^m = \text{argmax}_{s_0\dots s_T}\left(p(s_0\dots s_T,\ X_0\dots X_T|\hat{\Theta}_m)\right) \tag{2.17}$$

The empirical probability of being in state 2 at time t is then computed as $\frac{\#\{m,\hat{S}_t^m=2\}}{M}$.
Working on real data, state 2 (standing for "fishing") is attributed to the estimated state with the lowest mean for scalar speed, due to the fact that the vessel goes slower in that case.
The bootstrap is the most time consuming part of the estimation.

Simulated scenarios

The performance of the estimation method is assessed through simulations of trajectories based on various scenarios mimicking different levels of contrast in the movement and activity cha-

racteristics. The nine scenarios considered are all variants of a baseline scenario (S1, described hereafter). However they all share common properties on parameters values, chosen to fit the characteristics of the observed trajectories :

1) Movement does not privilege any turning direction, then we set $\eta_{r,1} = \eta_{r,2} = 0$, leading to an expected mean of V^r equal to 0 in both states ;
2) When cruising, vessel goes faster than when fishing. Then we set $\frac{\eta_{p,1}}{1-\mu_{p,1}} > \frac{\eta_{p,2}}{1-\mu_{p,2}}$: the expected mean of persistent speed is greater when steaming than when fishing ;
3) At a 15 minutes time step the time spent in each step is expected to be large in mean. Then, diagonal terms Π_{11} and Π_{22} of the transition matrix Π are large relative to the antidiagonal terms, meaning that the probability to stay in the same behavioral state is large relative to the probability to shift. This matrix is common to all scenarios and set to $\Pi = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$.

Trips are simulated following nine scenarios, with various degrees of mixture between states and with various number of observations (see Table 2.2 for detailed values).

Scenario 1 (baseline scenario). The contrast between $\eta_{p,1}$ (= 6) and $\eta_{p,2}$ (= 1) is large, as well as difference in autocorrelation parameters ("Steaming" state is uncorrelated while "Fishing" state is positively correlated).

Scenario 2-3 $\eta_{p,2}$ increases from 1 (scenario 1) to 2 (scenario 2) and 3 (scenario 3), resulting in an increase of the asymptotic expectation of V^p in state 2. Therefore the contrast in the expected asymptotic speed between state 1 and 2 decreases.

Scenario 4-5 $\mu_{p,2}$ increases from 0.5 (scenario 1) to 0.6 (scenario 4) and 0.8 (scenario 5), resulting in an increase of the asymptotic expectation of V^p in state 2. Therefore the contrast in the expected asymptotic speed between state 1 and 2 decreases. Moreover, the asymptotic variance of process V^p increase in state 2.

Scenario 6-7 In scenario 6, $\sigma^2_{p,1}$ and $\sigma^2_{r,1}$ increase from 1 and 0.5 (scenario 1) to 2 and 1 respectively, resulting to a higher asymptotic variance in state 1. In scenario 7, $\sigma^2_{p,2}$ and $\sigma^2_{r,2}$ increase from 0.5 and 0.1 (scenario 1) to 1 and 0.5 respectively, resulting to a higher asymptotic variance in state 2.

Scenario 8-9 The number of observations is shortened from 400 points (scenario 1) to 100 points in scenario 8 and 50 points in scenario 9. 400 and 50 points would represent respectively 100 and 12 hours data considered and were the maximal and the minimal length of trajectories considered.

2.2.3 Results

Results on simulated scenarios

Nine scenarios (corresponding to 9 set of parameters), mimicking different levels of contrast in the movement and activity characteristics, were simulated.

For each scenario, 100 trajectories are simulated, thus providing 100 parameter estimates and 100 estimations of the most credible sequence of behavioral states derived from the Viterbi

Scenario		η 1 2	μ 1 2	σ^2 1 2	$\mathbb{E}$ 1 2	$\mathbb{V}$ 1 2	n
1	p	$\begin{pmatrix} 6 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.5 \\ 0 & 0.2 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.1 \end{pmatrix}$	6 2	$\begin{pmatrix} 1 & 0.7 \\ 0.5 & 0.1 \end{pmatrix}$	400
	r				0 0		
2	p	$\begin{pmatrix} 6 & 2 \\ 0 & 0 \end{pmatrix}$			6 4		
	r				0 0		
3	p	$\begin{pmatrix} 6 & 3 \\ 0 & 0 \end{pmatrix}$			6 6		
	r				0 0		
4	p	$\begin{pmatrix} 6 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.6 \\ 0 & 0.2 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.1 \end{pmatrix}$	6 2.5	1 0.8	400
	r				0 0	0.5 0.1	
5	p		$\begin{pmatrix} 0 & 0.8 \\ 0 & 0.2 \end{pmatrix}$		6 5	1 1.4	
	r				0 0	0.5 0.1	
6	p	$\begin{pmatrix} 6 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.5 \\ 0 & 0.2 \end{pmatrix}$	$\begin{pmatrix} 2 & 0.5 \\ 1 & 0.1 \end{pmatrix}$	6 2	2 0.7	400
	r				0 0	1 0.1	
7	p			$\begin{pmatrix} 1 & 1 \\ 0.5 & 0.5 \end{pmatrix}$		1 1.3	
	r					0.5 0.5	
8	p	$\begin{pmatrix} 6 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.5 \\ 0 & 0.2 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.1 \end{pmatrix}$	6 2	$\begin{pmatrix} 1 & 0.7 \\ 0.5 & 0.1 \end{pmatrix}$	100
	r				0 0		
9	p						50
	r						

TABLE *2.2 – Parameters values for each simulation scenario. The matrix* Π *is identical for all scenarios. Asymptotic expectation and variance indicated are calculated from equations* (2.15) *and* (2.16)*, and rounded to first digit. They are asymptotic and must be considered as indicators of how the parameters affect the different processes.* n *is the number of observations along the simulated trajectories.*

algorithm.

Examples of velocity process obtained with parameters of scenarios 1, 3, 4 and 7 are represented on Figure 2.4. These scatter plots highlight the different degrees of mixture between the two states, depending on the scenario.
Knowing the true value of each parameter, estimation errors are computed and summarized using box plots (Figure 2.5). Results are shown only for process V^p, as trends are similar on process V^r (results for V^r are shown in appendix A.3). Moreover, as the true sequence of behavioural states is known, a misclassification rate is also computed and displayed using box plots (Figure 2.6).

Scenario 1-3 For all parameters, the width of the box plots increases from scenarios 1 to 3, revealing that decreasing the contrast between $\eta_{p,1}$ and $\eta_{p,2}$ has a negative impact on the estimation of all parameters (Figure 2.5). Moreover, the misclassification rate of the behavioral states is also increases. Even if it remains low for scenarios 1 and 2, it increases for scenario 3 (more than 50% of the 100 simulations result in a misclassification rate greater than 0.15, Figure 2.6). Looking at Figure 2.4, this large misclassification rate is explained by the large degree of mixture between states in scenario 3.

Scenario 1, 4-5 Increasing $\mu_{p,2}$ (while keeping $\eta_{p,1}$ and $\eta_{p,2}$ unchanged) increases estimation's uncertainty over level and autocorrelation parameters η and μ (Figure 2.5). The misclassification rate also increases, with a low increase for scenario 4, and a larger one for scenario 5 (Figure 2.6). Indeed, there is an increase in the degree of mixture between states from scenario 1 to scenarios 4 and 5 (see Figure 2.4 for scenario 4).

Scenario 1, 6-7 In scenario 6, increasing variance parameters $\sigma^2_{p,1}$ in state 1 increases slightly the uncertainty over the estimates of level and variance parameters $\eta_{p,1}$ and $\sigma^2_{p,1}$. The same effect is noticed in scenario 7 when variance parameters in state 2 increase (Figure 2.5). The misclassification rate remains stable between scenario 1 and 6, but increases for scenario 7 as the processes in both states have in this case the same variance parameters (Figure 2.6). Indeed, there is an increasing in the degree of mixture between states from scenario 1 to scenario 7 (Figure 2.4).

Scenario 1,8-9 When the number of observations is shortened, estimation's uncertainty increases for all parameters (Figure 2.5). Moreover, the misclassification rate is also impacted, getting worse as the observation's length gets shorter (Figure 2.6). Looking at estimates of $\hat{\Pi}_{22}$ (for instance, the same can happen for $\hat{\Pi}_{11}$), it is worth noting that considering only 50 data points as in scenario 9 can lead to estimate $\hat{\Pi}_{22}$ close to 0. This results in the identification of only one behavioral state, and then a large misclassification rate.

More generally, it is worth noting that for all scenarios, estimations are unbiased. Moreover, except for scenario 7 where variance parameters are equal in both states, the variance of estimators is greater for state 1 parameters than for state 2 parameters, as the variance parameter is larger in the first state ($\sigma^2_{p,1} > \sigma^2_{p,2}$).

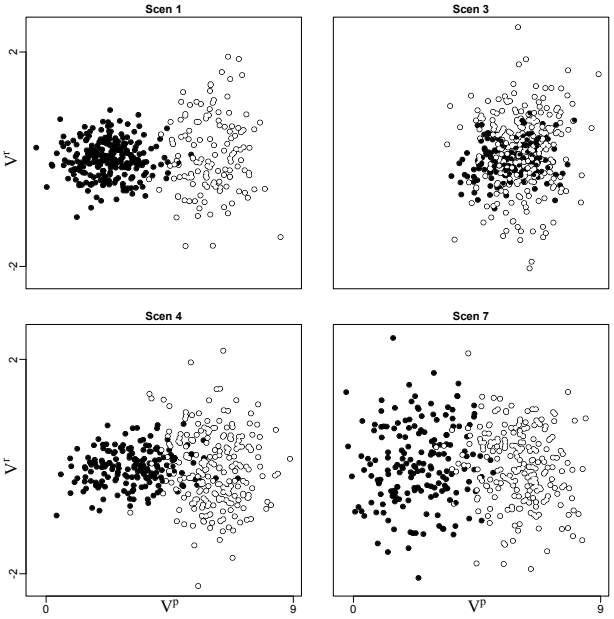


FIGURE *2.4 – Simulated velocity processes for scenarios 1, 3, 4, and 7 (see Table 2.2). The persistent speed V^p is represented along the x axis and the turning speed is represented along the y axis. Black dots are for fishing, white dots for steaming.*

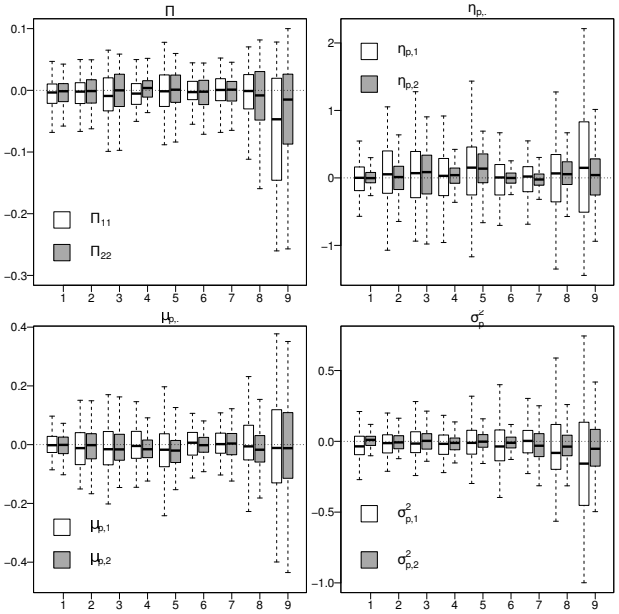


FIGURE *2.5 – Estimation errors (estimated value minus the true value of each parameter, on the y axis) obtained for the 9 simulation scenarios (x axis) presented in Table 2.2. Box plots represent the variability between the 100 simulations for each scenario. Only estimation errors for process V^p are presented, white and grey box plots are for parameters estimates in steaming and fishing respectively. The whiskers represent here at most 1.5 times the interquartile range. Outliers are not plotted.*

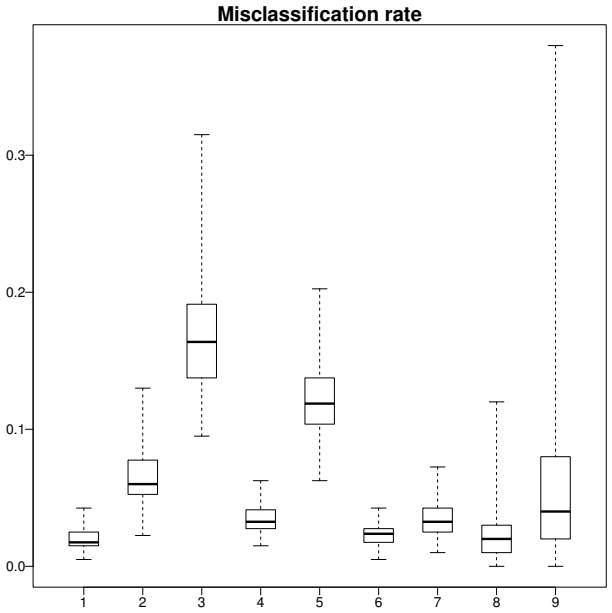


FIGURE *2.6 – Box plots of misclassification rate from the Viterbi algorithm (on the y axis) for the 9 simulation scenarios (on the x axis) presented in Table 2.2. Box plots represent the variability between the 100 simulations for each scenario. The whiskers height is at most 2 times the interquartile range. Outliers are not plotted.*

Results on RECOPESCA data

The four observed trajectories are represented on Figure 2.7 together with the estimated probabilities of fishing at eaxh observed position of the vessel. Velocities associated to these trajectories are represented using scatter plots on Figure 2.8. The uncertainty in distinguishing the behavior of fishing from steaming is low for all trips, as the estimated probability of fishing is most of the time 0 or 1 (see Figure 2.8). However, misidentification between fishing and steaming might occur at certain turning points (see Figure 2.7).

Estimations for the 14 parameters, the estimated proportion of time spent fishing and mean scalar speed in each behavioral state are presented on Figure 2.9. As expected, parameters Π_{11} and Π_{22} are high on those four trips showing a persistence to stay in a given activity. Parameters η_r are estimated quite close to 0, except on trip A. In this case, $\eta_{r,1}$ is slightly negative, stemming from the tendency of this vessel to always turn in the same direction during this trip.

The four trips have different patterns, trip A has an erratic fishing activity at low speed, trips B and C have a similar constant fishing activity pattern, with a constant speed and a steady course, and the trip D is a mix between strongly autocorrelated speed patterns and brutal changes.

For trip A, estimated steaming and fishing states are clearly separated on scatter plot of velocity (Figure 2.8), steaming being concentrated at high values for V^p (large value of $\eta_{p,1}$) and fishing being more dispersed at lower values of V^p (smaller value of $\eta_{p,2}$). Steaming represents 20% of the vessel's activity during the trip and takes place at high speeds (mean around 8.4 knots), while fishing represents 80% of the vessel's activity, it occurs at a lower speed (mean of 2.6 knots) and is more erratic.

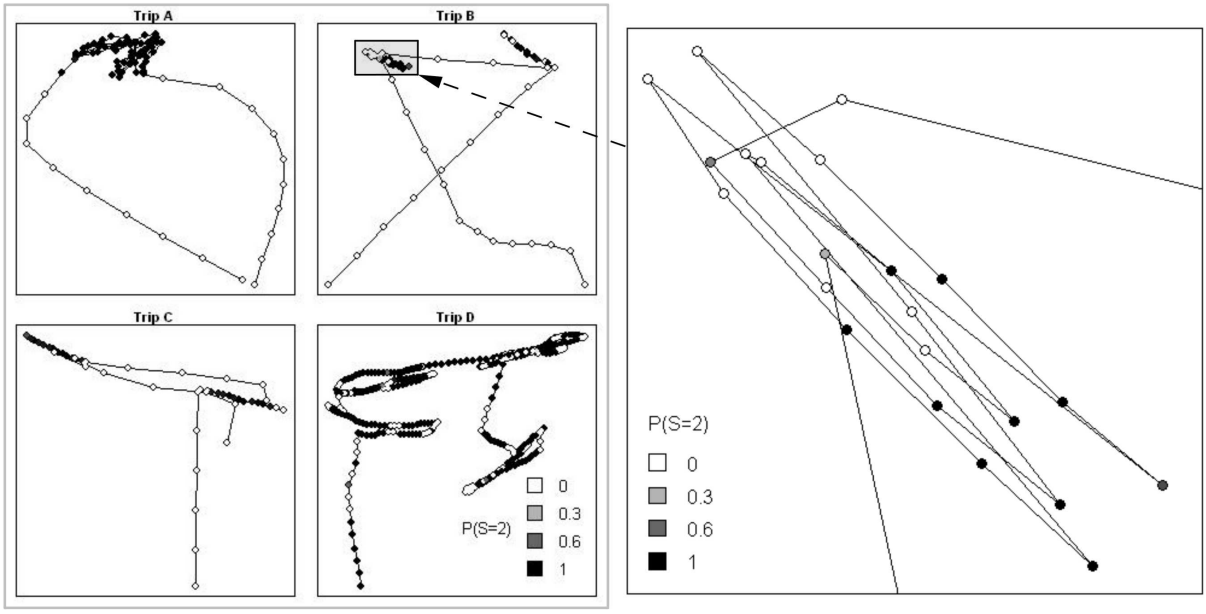


FIGURE *2.7 – The four trajectories observed from RECOPESCA position records. Estimated probabilities of being in state 2 (fishing) for each trip (A, B, C and D) is plotted, from 0 (white dots) to 1 (black dots). The right panel is a zoom over a specific zone of trip B (shaded rectangle). Estimations for trip D can not be interpreted as steaming/fishing (see text).*

For trips B and C both components of the velocity process, V^p and V^r, have a very low variability (Figure 2.8). In those two cases, parameters $\mu_{.,2}$ and $\sigma^2_{.,2}$ are small, resulting in a process (while fishing) with a small variance. Fishing then occurs at constant speed and result in a steady course. By contrast, the steaming state are estimated with large variance parameters ($\sigma^2_{.,1}$ have large values), then, as expected after simulation analysis, parameters estimates of V^p and V^r for this behavioral state are more uncertain.

For Trip D, estimated steaming and fishing states are more mixed. Parameters estimates show a low uncertainty for fishing and a larger uncertainty for steaming. Trip D has more than 400 observed positions, therefore, given the results of the simulation analysis, the uncertainty of state 1 estimated parameter can be associated to a large variance in the process (large values for $\sigma^2_{.,1}$). Moreover, state 2 is here characterized by a parameter $\mu_{p,2}$ really close to 1, traducing a highly autocorrelated V^p process in state 2. Interpretation of results obtained in trip D is questionable. Behaviors are mixed along the whole trajectory and seem unrealistic in terms of steaming/fishing (Figure 2.7). Figure 2.10 presents the scalar speed process for this particular trip. It shows that scalar speed does not discriminate the two states : both fishing and steaming present high speed values. However it is not realistic considering the vessel can operate an otter trawl at 10 knots. Actually, the 2 states are likely separated out based on the magnitude of autocorrelation in the V^p process, state 2 being highly autocorrelated ($\mu_{p,2} \approx 1$) with a small variance parameter (low $\sigma^2_{p,2}$), and state 1 being less autocorrelated and with a larger variance parameter (larger $\sigma_{p,1}$). On Figure 2.10, the autocorrelated process corresponds to a portion of sine waves, the other state is noise.

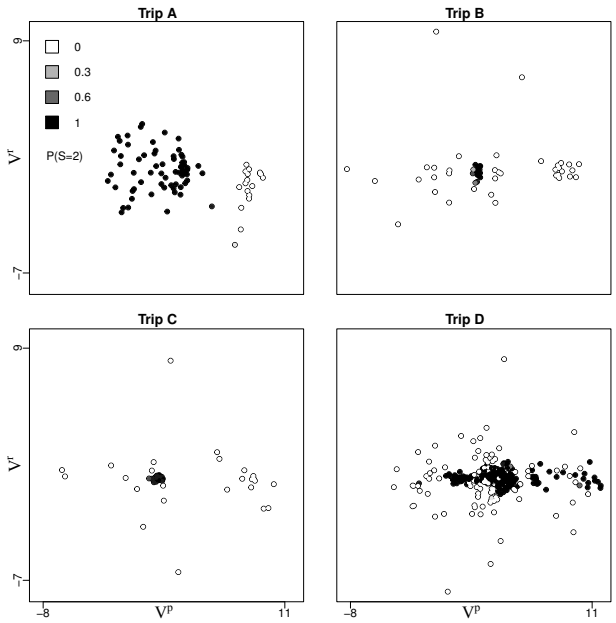


FIGURE *2.8 – Velocity process for each trip (A, B, C and D). The persistent speed V^p is represented along the x axis and the turning speed is represented along the y axis.. The estimated probability of being in state 2 (fishing) for each trip is plotted, from 0 (white dots) to 1 (black dots).*

2.2.4 Discussions and perspectives

This paper provides a first application of an AR process coupled with a hidden Markov chain to describe the movement of fishing vessels. The AR process allows a general framework with Gaussian properties and interpretable parameters. In this paper, it is shown how the velocity process viewed as an AR process can be used in order to analyse fishing vessels trajectories. Application to the RECOPESCA data set provides insights to the understanding of fishing vessel activities at a fine spatio-temporal scale. Results over the four studied vessel's highlight differences between-trajectories (different types of vessels and fishing activities, Biseau, 1998) and within-trajectories (between steaming and fishing) in terms of time series characteristics (means, variance, autocorrelation) that can also be translated in terms of physical patterns (fast, slow, erratic, steady). These results show that the fishing activity performed influence the estimates of the velocity process parameters.

The model considered here has two states, steaming and fishing, that could be similar to a "migrating"/"foraging" pattern adopted for animals (Jonsen *et al.*, 2007), whereas a three-states model can be used (Vermard *et al.*, 2010; Walker and Bez, 2010). This was made possible thanks to a pre-treatment of the data that consists in removing positions in port (that could be associated with a third with state V=0) but also because each studied fishing vessel operates with suitable gears that do not require research or stopping phase. If a two-states model is realistic here it could be more relevant in other cases to adopt a three or more states for trips during which several gears can be operated or several métiers can be performed. A model with "transition" states can also be adopted to deal with problems due to time step acquisition, and specifying different parameters for each fishery (Peel and Good, 2011). It is to note that

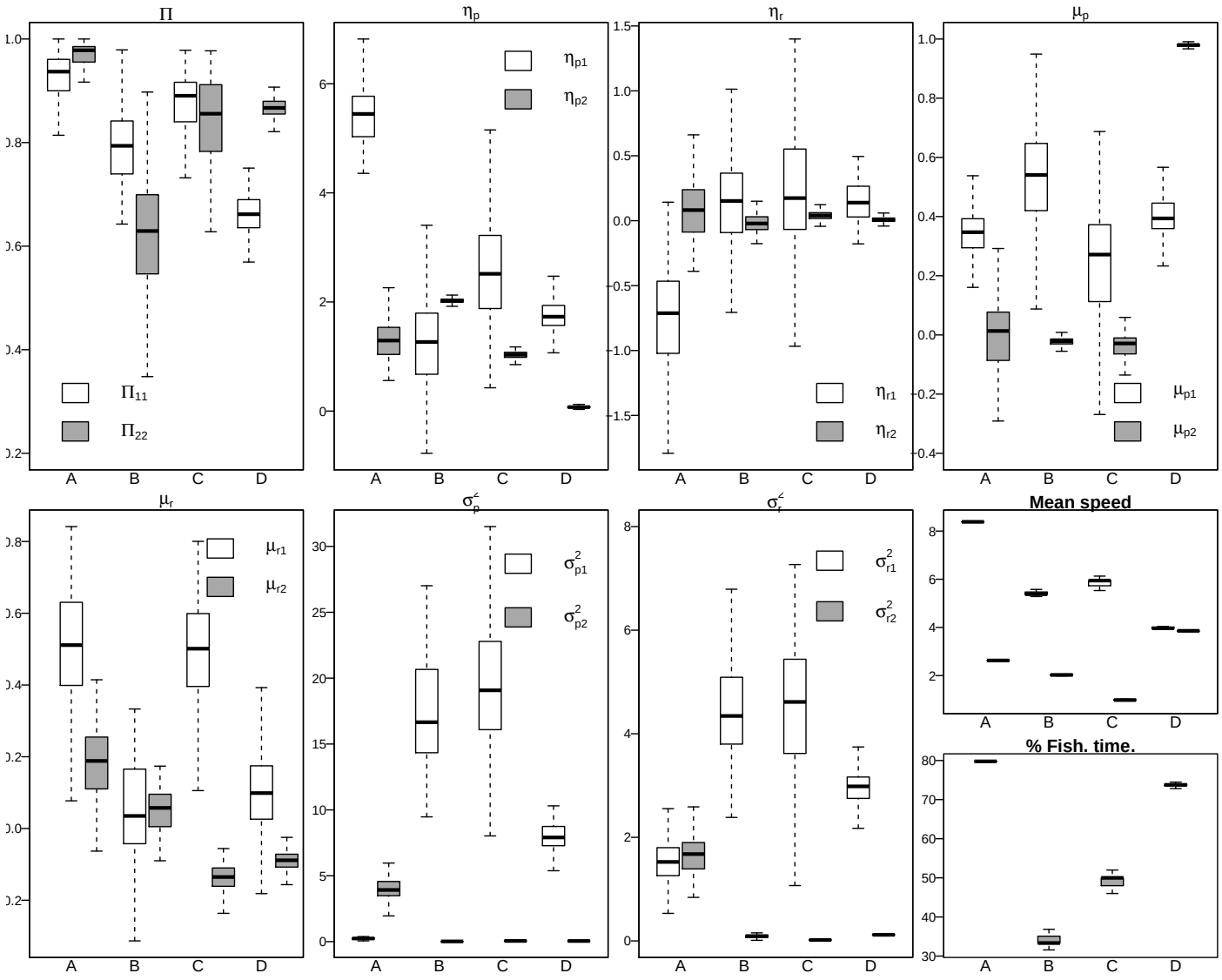


FIGURE *2.9 – Estimations for parameters in each trip. The mean speed in each state and the proportion of time spent fishing are also presented. Box plots represent the variability over estimates as assessed by the bootstrap procedure. White boxes stand for state 1, grey ones for state 2.*

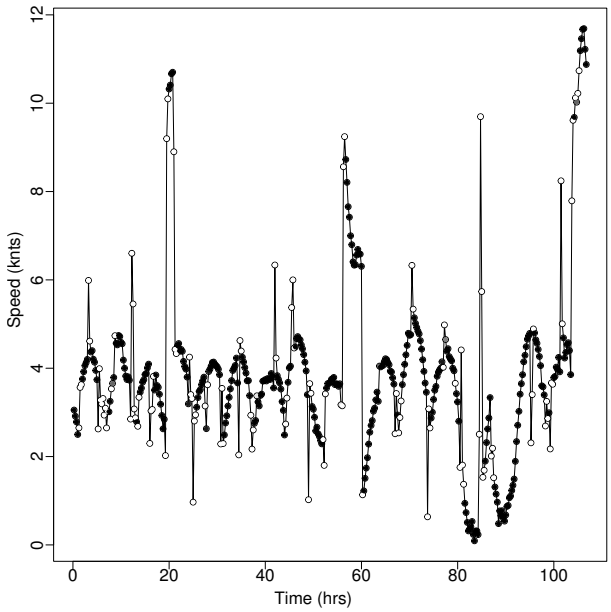


FIGURE *2.10 – Scalar speed process for trip D, with estimated states. Probability of being in state 2 is plotted from 0 (white dots) to 1 (black dots). The model seems to fail to attribute behavioral state with respect to the speed, as the scalar speed is high for both states. The state attribution here is based on autocorrelation (black dots and white dots for high and low autocorrelation, respectively).*

increasing the number a state would not add any difficulty to the method presented here (Jonsen *et al.*, 2013b). A more challenging alternative to these choices would be to consider a state space model where the number of states is a parameter to infer.

However, one has to question how to interpret the autocorrelation. In this study, the trip D (trip of a 22 meters bottom trawler) is disentangled into two states, one associated to highly autocorrelated persistent speed ($\mu_{p,2} \simeq 1$) and one associated to a less correlated persistent speed. It is to wonder whether the estimated autocorrelation is of interest for the user's purpose (here, knowing when the vessel is fishing) or whether it is the expression of an external factor that does not reveal the hidden behavior. In the case of the trip D presented here, the autocorrelated pattern might be the consequence of external factors that should be removed to accurately identify the steaming/fishing sequences. In that specific case, tide currents might be a cause of autocorrelated patterns. This covariate should then be included for future modeling of the vessels movement.

Along with this application on data, a simulation approach on nine contrasted scenarios is performed. The importance of the duration of observation is established as already noted by Vermard et al. (2000) : the longer the trajectory, the better the estimation. Then, we recommend that results are interpreted with caution when dealing with small vessels as they might not have long trips enough to reliable estimation. In order to apply this model to VMS mandatory data, this has to be taken in account has a potential problem. This model might not be well suited for VMS data as the linearity hypothesis would be even stronger at a 60 minutes time step, and the irregularity makes the use of an AR process inappropriate. Moreover, the simulation

approach pointed out problems to identify two behaviors when the contrast between them is too small. In practice, it implies that two "similar" fishing activities (for instance dredging and trawling) might not be distinguished).

The AR process presented here, that allows the introduction of autocorrelation in the velocity process, was used by Jonsen *et al.*, 2005 to model seals movement, but was never applied to fishermen's movement which are often compared to top predators movement (Bertrand *et al.*, 2007). This approach allows to model the bivariate velocity process (which fully describe the trajectory) with a unique Gaussian structure instead of two separated distributions as in previous models (Vermard *et al.*, 2010; Walker and Bez, 2010).

An interesting point of the model is its link with continuous time models. It is known that an AR process (with $\mu > 0$) is a discrete version of the continuous time Ornstein Ulhenbeck process (OUP), as long as the time step is regular (Johnson *et al.*, 2008). Indeed this model can be seen as an OUP sampled at discrete time coupled with a HMM, and there exists a bijection between the MLE of a positively correlated AR process and MLE of the regularly sampled OUP (see appendix A.4 for proof). The OUP is known to be the solution of a specific stochastic differential equation (SDE). SDEs are a general and challenging tool to model spatial trajectories (Brillinger, 2010), as its continuous time property allows to deal with irregularity in data but also to integrate spatially continuous covariates that rule the individual dynamics. However, it is to mention that SDEs would require to have instantenous velocities, and not averaged velocities as in the presented model. This is a line of research we wish to explore further to circumvent modelling difficulties of VMS mandatory data.

The parameter estimation is performed using the Baum Welch algorithm and the reconstruction of the hidden state sequence is achieved thanks to the Viterbi algorithm. In order to compare results, inferences were also performed in a Bayesian framework (not shown here) using MCMC methods, and estimates showed similar results. Actually, there are several techniques to estimate parameters in Hidden Markov Models (see Jonsen *et al.*, 2013b). Considering the AR process presented here, the Baum Welch algorithm does not need numerical techniques as the equations associated have analytic solutions. This allows to code the algorithm entirely without any toolbox, which is, in our opinion, an advantage, even if this might not be faster (Zucchini and MacDonald, 2009). The uncertainty about estimates is assessed using bootstrap methods instead of Fisher matrix, that could be given as an output of the algorithm. Bootstrap methods might be longer to compute (it's the most time-consuming part of the algorithm) but it does not rely on asymptotic assumptions which would not stand here for at least 3 of the 4 trips presented (Efron and Tibshirani, 1993; Zucchini and MacDonald, 2009).

2.3 La vitesse par rapport à la masse d'eau, un bon proxy pour les activités de pêche ?

La discussion de la section précédente fait l'état d'une forte autocorrélation dans la série des vitesses des navires, utilisées comme variables résumées de la trajectoire pour inférer les comportements. Cette forte autocorrélation dans le processus des vitesses semble être en partie responsable de la difficulté à segmenter la trajectoire sous la forme d'une séquence de comportements clairement identifiés comme "en pêche" ou "en route".
Les vitesses, sur lesquelles on s'appuie pour la segmentation, sont calculées depuis les positions GPS et sont donc directement calculées par rapport au sol. Hors, l'expertise halieutique suggère que pour nombre de navires de pêche (notamment les chalutiers), la vitesse de déplacement par rapport à la masse d'eau pourrait être un meilleur indicateur de l'activité de pêche que la vitesse par rapport au sol.
Cette section traite donc de la considération des courants dans la détermination de l'activité des navires de pêche : La détection des moments de pêche par les modèles existants est elle améliorée lorsque l'on utilise la vitesse par rapport à la masse d'eau ? Plus généralement, la considération des courants de surface nous apprend elle sur la dynamique des navires en Manche Est ? Ce travail a donné lieu à une soumission dans la revue Aquatic Living Ressources. Il a été écrit en collaboration avec Mathieu Woillez, Stéphanie Mahévas, Youen Vermard et Etienne Rivot.
L'auteur présente à nouveau ses excuses au lecteurs pour le passage à l'anglais.

Is the speed in relation to the water mass a better proxy for fishing activities than speed in relation to the ground?

Pierre Gloaguen[a]*, Mathieu Woillez[b], Stéphanie Mahévas[a], Youen Vermard[a], Etienne Rivot[c].

2.3.1 Introduction

Understanding the fishing vessel dynamics at a fine spatial scale is of great interest to define appropriate management measures. Fishing trips typically consist in a sequence of different activities like steaming and trawling. For trawlers, identifying and characterizing trawling sequences during a fishing trip is a key step to understand the spatio-temporal dynamics of fishing effort (Vermard *et al.*, 2010) or to characterize the costs relative to trawling and then to evaluate the economic performance of fisheries (Pelletier *et al.*, 2009).
The analysis of sequence of GPS positions from Satellite-based Vessel Monitoring System (VMS) has received a growing attention in the last decade (Mills *et al.*, 2007; Hintzen *et al.*, 2010; Rijnsdorp *et al.*, 2011; Russo *et al.*, 2011a). Various models have been developed to analyse fishing activity from VMS data. While mathematical and statistical methods differ, all rely somehow

on the idea that the speed in relation to the ground, derived from successive GPS positions, provides critical information to infer the fishing vessel activity. The simplest methods consider that fishing occurs when the vessel speed is below a particular threshold (Berthou *et al.*, 2013). Hidden Markov Models that consider the sequence of fishing behavior as a Markov process (Vermard *et al.*, 2010; Walker and Bez, 2010; Peel and Good, 2011; Gloaguen *et al.*, 2015) have also performed well to capture realistic sequences of fishing/non fishing behaviors.
However, although fishing vessel behavior and the operating mode of fishing gears are directly impacted by the movement of the water mass, no previous study has investigated the importance of considering the influence of currents on the fishing activity. Surface currents can be very strong and fluctuate in some coastal fisheries (e.g. in the Eastern Channel), and may disconnect the speed in relation to the ground from the speed in relation to the water mass. For instance, in a recent paper, analyzing trips of otter trawlers operating in the Eastern Channel, Gloaguen *et al.* (2015) have shown oscillations with approximately twelve-hour period in the speed in relation to the ground , thus suggesting that the speed in relation to the ground rather reflects surface currents than fishing activity. Also, both the behavior of benthic or pelagic fish, and the efficiency of some fishing gear are likely to interact with currents.
In this paper, we investigate the interest of considering the surface currents to analyse fishing activity from VMS data. The approach is illustrated with data from fishing trips of trawlers in the Eastern Channel where the surface currents are known to be particularly strong, and are suspected to structure part of the fishing activity. We propose a simple approach to combine VMS data with surface currents derived from circulation models to estimate the speed in relation to the water mass along the trajectories. Then, using three previously published models, a simple speed threshold and two others based on Hidden Markov Models, we investigate how considering the speed in relation to the water mass instead of in relation to the ground may improve inferences on fishing activity and potentially improves our capacity to identify sequence of fishing behaviors.

2.3.2 Material and methods

2.3.3 GPS positions from RECOPESCA data

We used VMS data from the RECOPESCA research project, a voluntary based project developed by Ifremer to improve the assessment of the spatial distribution of catches and fishing (Leblond *et al.*, 2010). GPS positions of a sample of French volunteer fishing vessels are recorded with a short and highly regular time step ($\simeq$ 15mn). Five vessels operating in the Eastern Channel were considered. These vessels differ in terms of length and engine power. They all use bottom towed gears, which were used by 83% of French offshore fishing vessels larger than 12 meters (thus, mandatory equiped with VMS) performing in the channel (figures for 2012, from Leblond *et al.*, 2014b) . Moreover, one of them uses trammel nets (see table 2.3). Overall, 1668 fishing trips were recorded between 2007 and 2011, and used for this study.
To highlight results of our method at a trip scale, we chose a representative trip of the vessel n°

Vessel's ID	Length (m)	Engine Power (kW)	Fishing gears
1	10,30	106	Otter trawl, Dredges
2	12,00	162	Dredges, Otter trawl, Beamer Trawl
3	13,25	242	Trammel nets, Dredges, Beamer Trawl
4	21,00	442	Otter Trawl
5	22,50	371	Otter Trawl

TABLE *2.3 – Description of the 5 RECOPESCA vessels performing in the Eastern Channel. For each vessel, gears used are presented from the most used on the left to the least used on the right.*

5. This trip consists in more than 400 GPS positions recorded in the Eastern Channel, and the only fishing gear used during the trip was otter trawl. The vessel n° 5 is the largest vessel among our sample, and it is the one that goes the furthest from coasts. Also, an onboard observer from the ObsMer project (Dubé *et al.*, 2012) was present on this trip. Onboard observers bring critical additional information about fishing operations. The hauling time is precisely registered at the end of each trawling sequence. This information was then used to validate the fishing/non fishing segmentation of trips inferred from the different models. Moreover, for each fishing operation, catches are recorded.

Computing the speed in relation to the water mass

For each trip, a sequence of speed in relation to the ground is first derived from the sequence of recorded GPS positions available at each time step, and then combined to a surface currents field derived from a circulation model to estimate speed in relation to the water mass at each time step. The main lines of the method are given below. Technical details are given in appendix.

For each trip, the observation first consists in a sequence of GPS positions (latitude and longitude) denoted $(Z_i)_{i=0...n}$ recorded at times $t_0 \dots t_n$ that are easily converted into a sequence of two dimensional coordinates on a regular distance grid denoted $(X_i)_{i=0...n}$. Then, the sequence of vectorial speed in relation to the ground denoted by $(V_i^{raw})_{i=0...n-1}$, is derived from the sequence of positions assuming a linear path between points (see appendix). We denote by $\| V_i^{raw} \|$ the scalar norm of the vectorial speed in relation to the ground.

The speed in relation to the ground mostly results from the addition of forces due to the engine and to the surface currents. The surface currents component can be removed to obtain the speed in relation to the water mass, hereafter denoted by V^{wtc} (vectorial) and $\| V_i^{wtc} \|$ (scalar norm). Surface currents are directly derived from outputs of the hydrodynamical model MARS 3D developed by Ifremer (Lazure and Dumas, 2008) to derive operational prediction in coastal oceanography (Lecornu and De Roeck, 2009). Because the circulation model provides surface currents on a 4km× 4km grid every hour, an interpolation is used to associate a model output of surface currents to each recorded position. The speed in relation to the water mass is then obtained by removing the current component from the speed in relation to the ground (see appendix B.1).

Identifying sequence of trawling from the speed in relation to the ground or to the water mass

For each fishing trip, the sequence of speed in relation to the ground $(V^{raw})_{i\geq 0}$ or in relation to the water mass $(V^{wtc})_{i\geq 0}$ is used to draw inference on the fishing activity. Three previously published models where used. Because no new model development was done in this work, we briefly summarize the principles of those models.

- Speed threshold model (Berthou *et al.*, 2013). The speed threshold method considers only two different possible behaviors : trawling and steaming. At each time step i in the sequence $i = 0 \dots n-1$, the model considers that the fishing vessel is trawling if $\| V_i \| < 4.5$ knots, and steaming if $\| V_i \| > 4.5$ knots.
- Bayesian Hidden Markov Models (Vermard *et al.*, 2010). The model considers two possible behaviors, trawling and steaming, and assumes the sequence of hidden behaviors $(S)_{i=0\dots n-1}$ follows an homogeneous Markov process. The movement is considered linear between two time steps and is described by the scalar speed $\| V_i \|$ and the turning angle θ_i between two successive time step. The speed and turning angles are drawn in a prior distribution with parameters specific to the behavior, which provides a way to estimate the behavior from the sequence of speed and turning angles. The model is fitted using MCMC simulations. The model is hereafter designed as the correlated random walk (or CRW) model.
- Hidden Markov Models with autoregressive speed process (Gloaguen *et al.*, 2015). The inspiration of the third model is quite similar to the Hidden Markov Model developed by (Vermard *et al.*, 2010). The sequence of hidden behaviors $S_{i=0\dots n-1}$ also follows an homogeneous Markov process, but the sequence of speed is supposed to follow a first order autoregressive process to capture an inertia in the movement. The estimation procedure has been developed in the maximum likelihood framework using an Expectation-Maximization algorithm. This model is hereafter designed by the autoregressive (or AR) model.

The general framework of state space models applied to fishing vessel activities is recalled on Figure 2.11.

The performance of the three models was measured through the capacity to identify the sequence of true behavior (trawling or steaming). At each time step i, the behavior inferred from the model was compared to the true behavior derived by the onboard observer. Reliable series of onboard observer data was only available for one representative fishing trip of the vessel n° 5. The misclassification rate was calculated as the percentage of time steps where the behavior has been misidentified calculated over the 400 time steps of the trip.

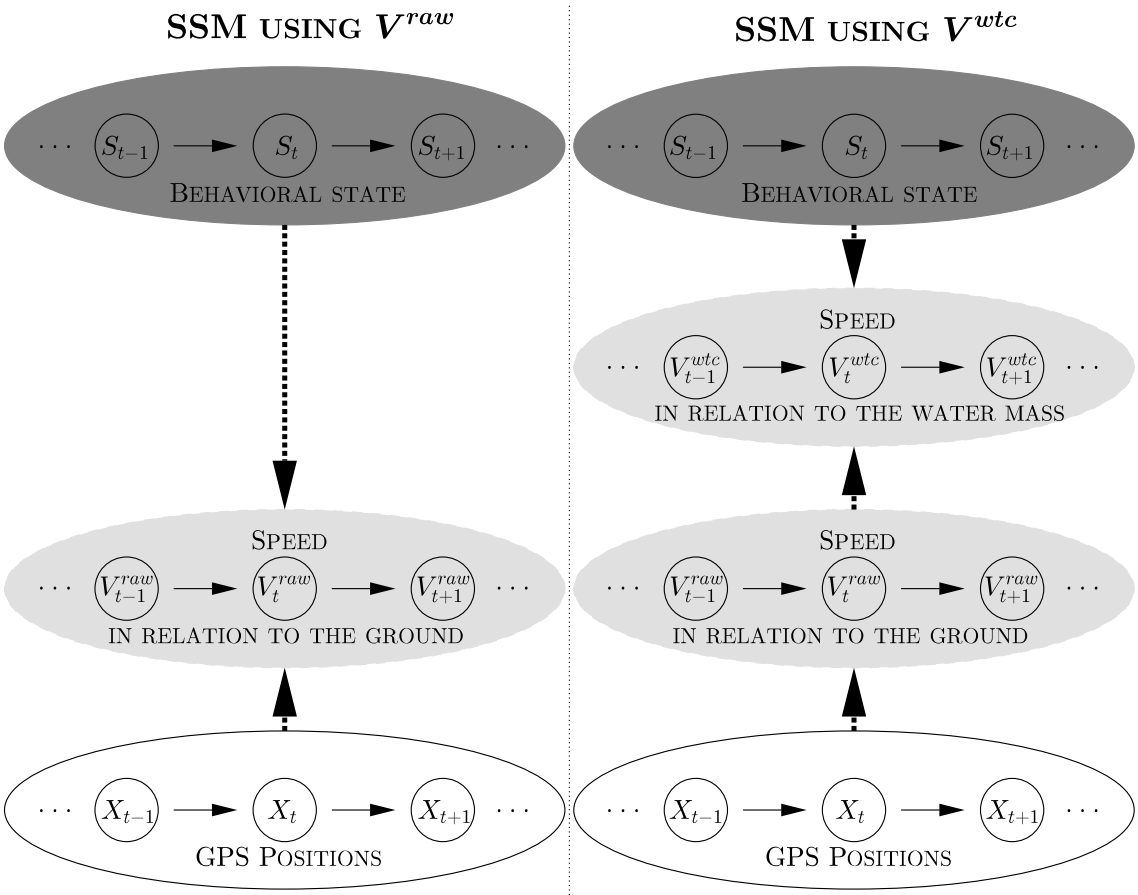


FIGURE *2.11 – Formal state space modeling to infer fishing vessel activity. White circles are observed, grey circles are computed, and black circles are hidden, and need to be estimated.*

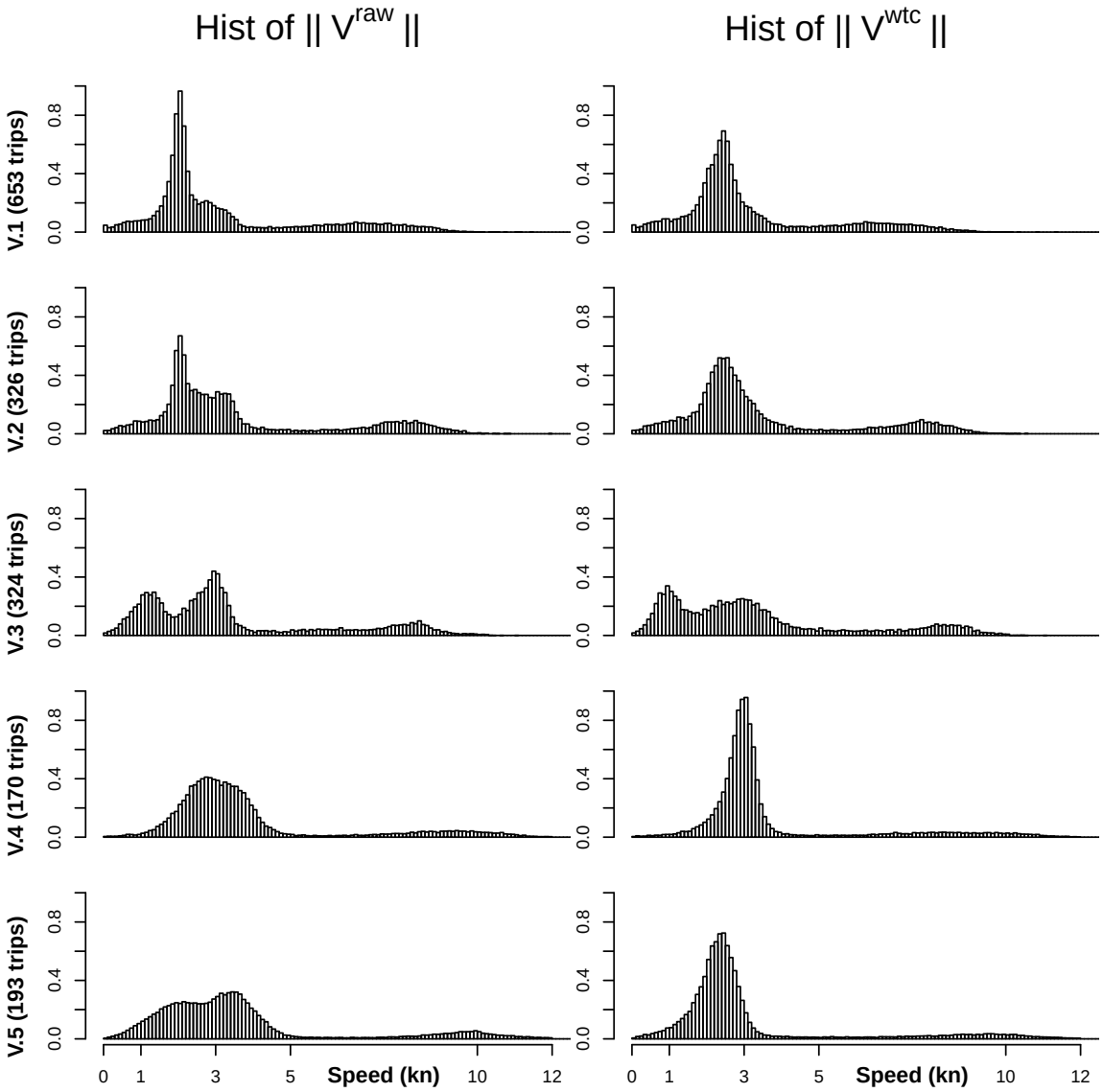


FIGURE *2.12 – Influence of surface currents on speeds distribution for each vessel (1 line per vessel). On the left are speed distributions before the withdrawal of surface currents. On the right are the associated distributions after the withdrawal of surface currents.*

2.3.4 Results

Speed in relation to the water mass is more constant than speed in relation to the ground

For all vessels, most of the speeds $\| V^{raw} \|$ and $\| V^{wtc} \|$ are between 1 and 4.5 knots, and histograms exhibit a heavy tail at high speeds around $9 - 11$ knots (Figure 2.12). Overall, comparing the speed in relation to the ground and the speed in relation to the water mass along the fishing trips shows that the speed in relation to the water mass is less variable (Figure 2.12) than the speed in relation to the ground.

The reduction in the variance is more marked for the speed below 4.5 knots than for the high speed. For vessels n° 1 and 2, the histogram of speeds in relation to the ground below 4.5 knots exhibit two modes around 2 and 3.5 knots, whereas the speed in relation to the water mass has

Model used	Threshold	CRW Model, 2 states	AR Model
Results on V^{obs}	13.4	13	38.2
Results on V^{wtc}	13.8	14.7	23.7

TABLE *2.4 – Comparing misclassification rate for Fishing/steaming on 3 different models, whether the surface currents are removed or not.*

only one mode at about 2.5 knots. While histograms of the speed in relation to the ground of vessels n° 4 and 5 do not really exhibit two modes, the distribution of speed in relation to the water mass is much more concentrated around 3 and 2.5 knots, respectively. For fishing vessel n° 3, both distributions of speed in relation to the ground and to the water mass below 4.5 knots are bimodal.

Looking at the sequences of speed along the trips shows different types of sequences for speed below 4.5 knots and above 4.5 knots. While the sequence of speed higher than 4.5 knots rarely exceed 4 time steps (i.e. rarely exceed 1 hour), the sequence of speed below 4.5 knots can be much longer. When looking at speed below 4.5 knots, speed in relation to the ground exhibits sinusoidal variations between 2 and 4 knots with an approximate time period of 12 hours (Figure 2.13), while those fluctuations are dampened around 2.5 knots when looking at the speed in relation to the water mass.

For all vessels, the orientation at low speed (below 4.5 knots) in relation to the surface currents is easily computed. The focus is made on 4 general orientations, between 0 and 45 degrees (surface currents come from behind the vessel), between 45 and 90 degrees, between 90 and 135 degrees (in these two cases, surface currents are transverse to the vessel) and between 135 and 180 degrees (the vessel is facing surface currents). The time spent parallel to surface currents (going with or facing) is always above 60%, reaching more than 80% for vessel n° 5 (Figure 2.14). The heading of the vessel at low speed appears strongly correlated to the direction of the surface currents. In the Eastern Channel, main currents are mainly parallel to the coast, therefore, hauling in this region would be more likely in this direction.

Evaluating influence on fishing effort estimation

Performances of the 3 different models to segment a vessel trip between fishing and non fishing activity were tested using the speed in relation to the ground or to the water mass as the movement characteristics directly related to the behavior (fishing / non fishing). The misclassification rate obtained by comparing the observed behavior (from an onboard observer) with the estimated behavior is shown on table 2.4. Surprisingly, using the speed in relation to the water mass instead of in relation to the ground does not improve the classification performance, except for the HMM model with autoregressive process on the speed. However, overall, the misclassification rate remains higher for the HMM with an autoregressive process on the speed.

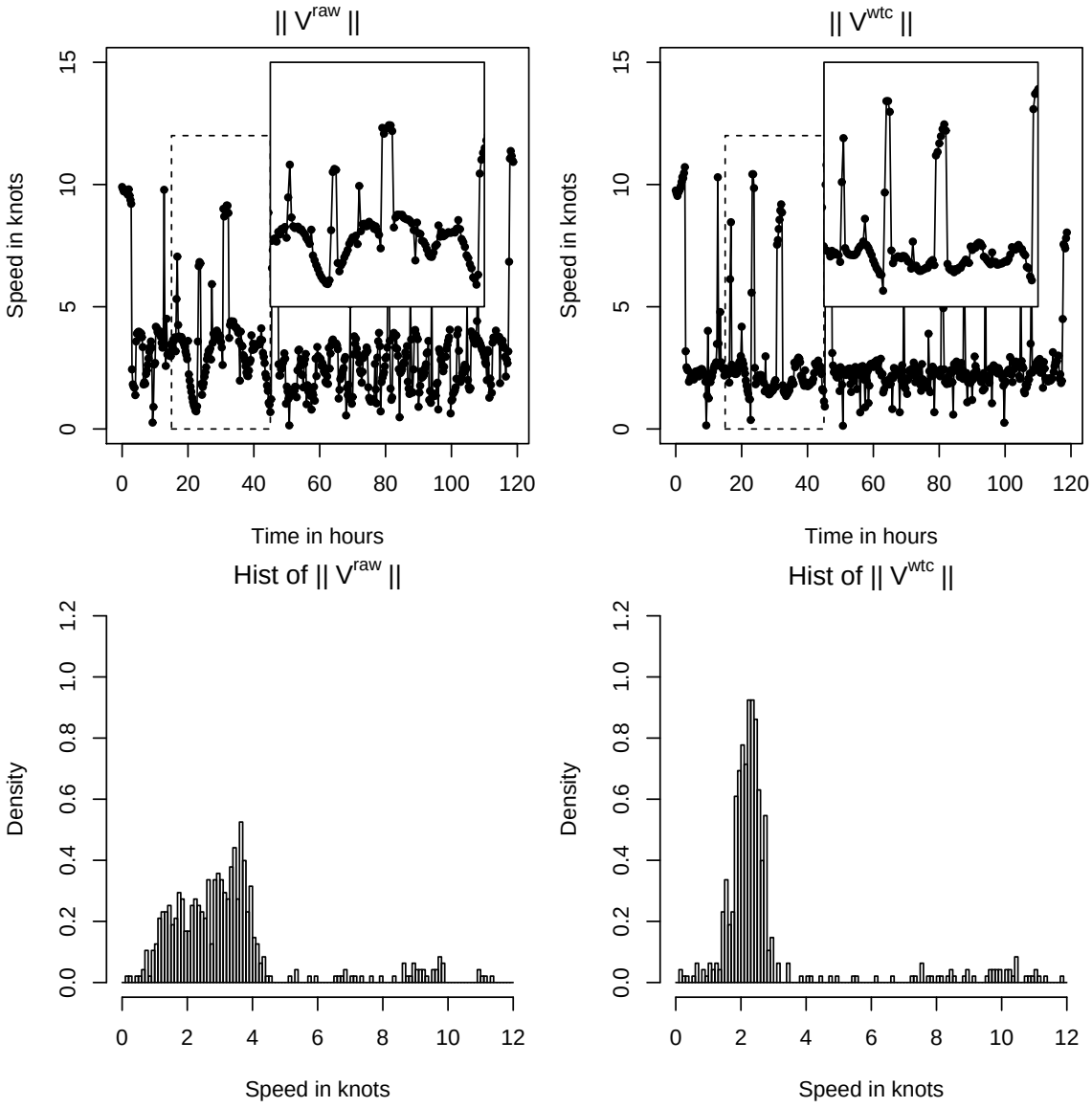


FIGURE *2.13 – Influence of surface currents on speed process for a given trip.* $\|V^{raw}\|$ *(left) and* $\| V^{wtc} \|$ *(right) are represented. Top line : Plotting of speed processes, a zoom is made on an oscillating pattern due to surface currents (on* $\| V^{raw} \|$*) that is flattened when surface currents are removed. Bottom line : Overall distribution of speeds for this trip are presented for each process.*

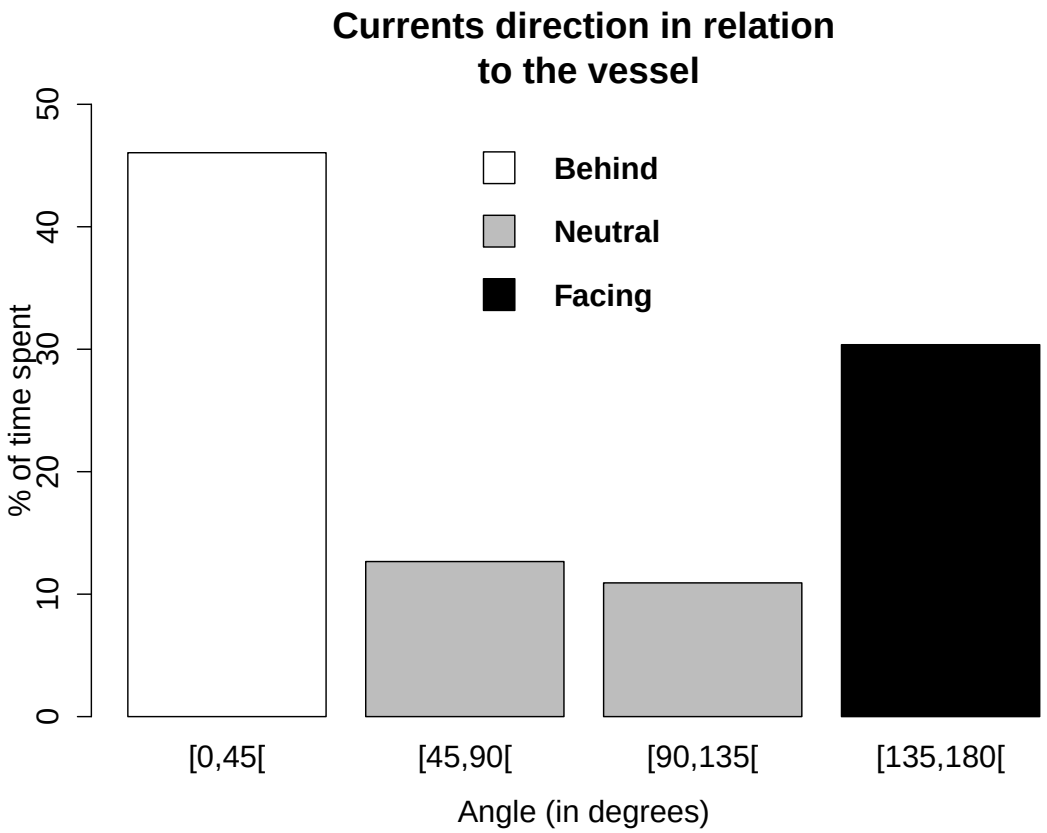


FIGURE *2.14 – Proportion of time spent by vessel n° 5 for different orientations in relation to surface currents. These orientations are considered when the vessel is at low speed (< 4.5 knots). There is one class each 45 degree. The left class is when the vessel has surface currents at its back. The two center classes is when the orientation in relation to surface currents is neutral, and the right class is when the vessel is facing surface currents.*

2.3.5 Discussion

Results suggest that the surface currents may highly structure the fishing activity and that it should be considered in models aiming at analysing fishing activity from VMS data. Overall, when considered together, those results strongly suggest that the higher variations of speed in relation to the ground between 2 and 4.5 knots, likely associated with trawling, could be due to the surface currents, while the speed in relation to the water mass remains more constant about 2.5 knots. Combined with the fact that fishing vessels spent most of the time in the same directions as the currents, this suggest that trawling activity is mainly performed in a direction parallel to the currents, and with a rather constant engine regime, leading to a rather constant speed in relation to the water mass while the speed in relation to the grounds varies with the currents.

As an important result, when considering speed below 4.5 knots preferentially associated to trawling behavior, our study revealed that the speed in relation to the water mass are much less variable than the speed in relation to the ground. This has an impact for HMM estimates, as the variance of the speed associated to each behavior should be lower. An important hypothesis of HMM developed by Vermard *et al.* (2010) and Peel and Good (2011) is the normality of speed distribution in a given behavior. It is shown here that the speed in relation to the water mass better supports this assumption than the usual speed in relation to the ground.

This study highlights that two modes at low speed in the raw speed process do not necessarily imply two different fishing behaviour (for vessels n° 4 and 5 for instance, Figure 2.12). However, if these two modes persist after the withdrawal (vessel n° 3 for instance, Figure 2.12), this might imply two different fishing behaviours in the process. In this case, the structure of state space models presented here, having only one fishing behavior, might not be well suited. Moreover, considering speed in relation to the water mass might be of great importance for economic studies. Indeed, exploitation costs strongly depends upon fuel consumption. Because the speed in relation to the water mass is a good proxy of the engine regime, it would be worth considering to use the speed in relation to the water mass instead of in relation to the ground to infer exploitation costs.

The orientation of fishing vessels in relation to surface currents shows that, at low speed, vessels spend most of their time (up to 80%) with or against the current. This suggest that surface currents orientations are an environmental driver of vessels dynamics in the Eastern Channel and should be taken into account when studying their displacements.

The importance of removing surface currents to segment a fishing trip between fishing and non fishing behaviour was also tested, using 3 different estimation methods and validation data from the Obsmer program. Surprisingly, using the speed in relation to the water mass instead of in relation to the ground do not improve the classification performance, except for the HMM

model with autoregressive process on the speed developed in Gloaguen *et al.* (2015). For the threshold method, the performance does not improve when using $\| V^{wtc} \|$ instead of $\| V^{raw} \|$, which is not surprising, as raw speeds below 4.5 knots remain below this limit when the surface current forces are removed (see Figures 2.13 and 2.12). However, surface currents must be taken in account to define the threshold value. In a zone having strong surface currents, this value should be higher than in a zone where surface currents are weak. For the first HMM model developed by Vermard *et al.* (2010), classification performance is not improved neither, as the mean speed when fishing does not change when using $\| V^{wtc} \|$instead of $\| V^{raw} \|$. For the model with an autoregressive process on the speed, the classification performance is improved, which was expected, as the withdrawal of currents reduce the autocorrelation in the process $\| V^{wtc} \|$, (Gloaguen *et al.*, 2015). Indeed, Gloaguen *et al.* (2015) showed that this model encountered some failures on vessels performing in the Eastern Channel. Instead of splitting a trajectory into fishing/non fishing segments, the segmentation was made on correlated/uncorrelated patterns. The authors suggested that those weaknesses of the model were due to oscillating patterns in the speed process. However, the misclassification rate of this model remains high. This suggests that, compared to the threshold and the correlated random walk model, this model might not be well suited to segment between fishing and non fishing activity during a trip.

This study was performed using vessels from RECOPESCA project using bottom gears (mainly otter trawls), therefore, conclusions might hold only for those type of gears, performing in regions having strong currents. The limited number of vessels (5) is a small sample regarding the amount of vessels performing in the Channel. However, the large number of fishing trips sampled (over 4 years) suggest that these results would be representative of a larger pattern.

As a recommendation, we suggest that surface currents be considered when studying speed processes of vessels (or fishes). We also suggest considering the main direction of hauling, given by surface current directions, when analysing fishing vessel dynamics. Finally, we suggest taking surface currents into account when defining a threshold for fishing/non fishing segmentation.

2.4 Discussion générale sur les HMM pour détecter l'activité de pêche

> *« Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. A first, though at first sight, not a very helpful principle, is that all models are wrong ; some, though, are more useful than others and we should seek those. [...] A second principle [...] is not to fall in love with one model to the exclusion of alternatives. »*
>
> **McCullagh and Nelder, 1989**

Les deux sections précédentes ont montré le développement d'un modèle HMM pour détecter l'activité des navires de pêche. Ce développement a donné lieu à 3 nouveautés par rapport aux approches HMM existantes en halieutique

1. Considérer le couple (V^p, V^r) [11] (voir les équations (2.11) et (2.12)) au lieu du couple vitesse et changement de direction, comme série temporelle susceptible de distinguer au mieux l'activité de pêche ;
2. Intégrer un terme d'autocorrélation dans ces variables, en considérant que celles-ci suivaient un processus autoregressif gaussien d'ordre 1 ;
3. Considérer les HMM en utilisant la vitesse par rapport à la masse d'eau, et non plus par rapport au sol.

Ces choix ont permis de s'interroger sur plusieurs hypothèses de la modélisation.

2.4.1 Quelles séries temporelles pour détecter l'activité ?

Les points 1. et 2. ci dessus ne sont pas déconnectés dans leur approche. En effet, dans l'optique d'introduire de l'autocorrélation dans les processus, pour pallier le problème de l'hypothèse d'indépendance des données, le processus AR(1) Gaussien est le plus répandu et intuitif. Or, de par leur structure, les processus des vitesses et des changements de direction utilisés jusqu'alors ne sont pas adaptés à cette modélisation. La vitesse scalaire est une grandeur positive, parfois proche de 0 dans notre cas, et le changement de direction est une grandeur circulaire, le support d'une loi Gaussienne n'est donc pas adapté pour décrire ces grandeurs. En revanche, la combinaison de ces deux grandeurs, telle que proposée avec V^p et V^r, résulte en deux processus aux supports réels. De plus, ces deux processus restent interprétables dans la problématique donnée, en termes de vitesses et d'ératicité de la trajectoire. D'un point de vue pratique, au vu du cas d'étude présenté ici (des chalutiers de fond), les processus de vitesse de persistance et de vitesse scalaire sont presque confondus. En effet, le déplacement des navires se fait globalement en ligne droite, que ce soit pendant la phase de déplacement vers la zone de pêche, ou la phase de chalutage. De notre point de vue, ces deux quantités gardent les propriétés qu'apportent la vitesse et les changements de direction (informations sur la vitesse et sur le caractère erratique),

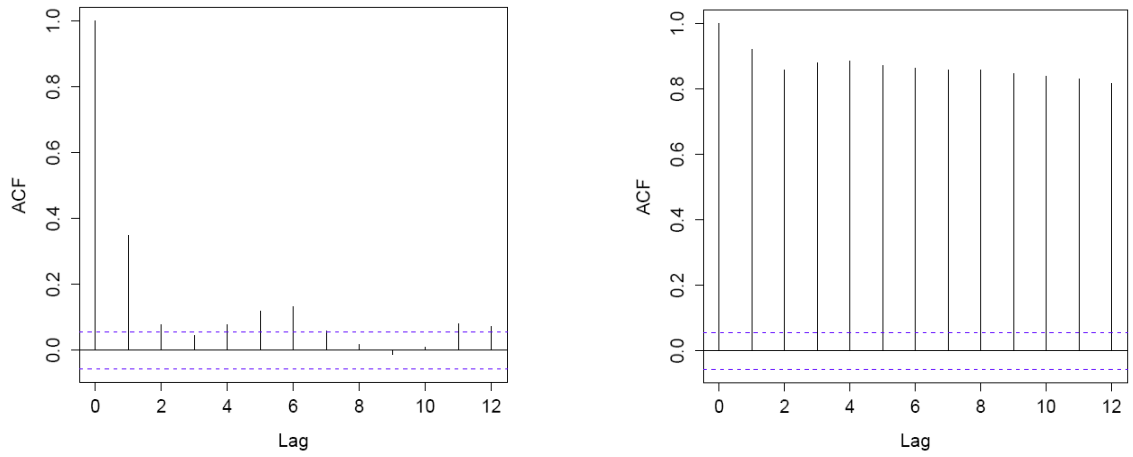
11. Dans la suite, on s'y réfèrera parfois par "Vitesse de persistance" et "Vitesse de déviation"

mais sur un support réel plus adéquat à la modélisation Gaussienne. Ainsi, nous pensons que ces deux variables sont aussi adaptées que les variables utilisées jusqu'alors pour détecter l'activité, avec l'avantage d'offrir un support plus connu et mieux maîtrisé, qui est celui des processus Gaussiens.
De plus, ce travail a permis d'aborder la pertinence de considérer la vitesse par rapport à la masse d'eau, et non plus par rapport au sol, comme variable résumée de la trajectoire sur laquelle s'appuyer pour inférer les comportements. Cette question est naturelle quand on considère un individu en milieu marin, où le support du déplacement est donc la masse d'eau. Ce choix de référentiel, plutôt que le sol, peut donc paraître légitimement plus adapté. Dans le cas présenté ici, travailler sur cette variable permet d'améliorer les performances du modèle AR HMM, car la vitesse par rapport à la masse d'eau est moins autocorrélée que celle par rapport au sol. La connaissance de cette vitesse a été rendue possible grâce aux modèles de circulation océaniques. Il convient de noter que cette option n'est pas toujours disponible, peut être coûteuse en temps, et nécessite une interpolation entre les sorties de modèles et les observations. Cet aspect est à prendre en compte, et ne favorise pas nécessairement l'utilisation de la vitesse par rapport à la masse d'eau.
De ce point de vue il convient de souligner que ces modèles se basent sur des quantités calculées, et non observées, la vitesse moyenne et les changements de direction moyens. Ces quantités sont toujours calculées, le plus souvent par interpolation linéaire. Et même si d'autres méthodes ont été proposées (Hintzen *et al.*, 2010), on se base toujours sur un processus approché pour détecter l'activité. Formuler un modèle dépendant seulement des positions permettrait d'éviter une telle approximation, et, potentiellement, de perdre la nature spatiale de la trajectoire. Cet aspect est en effet complètement omis en travaillant sur la vitesse et les changements de direction, on résume la trajectoire à de simples séries temporelles.

2.4.2 Quel modèle pour le déplacement conditionnellement aux comportements ?

Dans le modèle conditionnellement aux états, on a intégré un paramètre d'autocorrélation pour les séries considérées. Il convient de s'interroger sur la performance [12] de ce modèle par rapport aux modèles existants, supposant une indépendance conditionnelle des séries [13]. Au vu des résultats, il semble que ce modèle ait des performances moindres que le modèle proposé par Vermard *et al.* (2010). Ceci étant, il est intéressant de constater qu'un modèle plus général (et plus paramétré) ne permet pas de mieux répondre à la question de la détection d'activité. Il a été montré que l'hypothèse de non autocorrélation dépend du pas de temps de l'échantillonnage, et est d'autant plus discutable que le pas d'acquisition est court (Bez *et al.*, *in prep*). Les processus de vitesse montrés dans la section précédente ne satisfont clairement pas cette hypothèse (figure 2.15). Ceci étant, dans l'optique de détecter l'activité de pêche, négliger cette réalité ne semble pas réellement préjudiciable. Faire la distinction du comportement sur la base de

12. En terme de détection de l'activité
13. i.e. $Y_t|S_t \perp Y_{t-1}$



(a) Pêche *(b) Non Pêche*

FIGURE *2.15 – Exemple d'autocorrélation empirique du processus V^p dans les différentes activités. Le pas de temps d'acquisition est ici de 15 minutes. Un lag correspond à un pas de temps. Supposer l'indépendance de V_t^p et V_{t+1}^p néglige donc une corrélation significative. La ligne bleue représente le seuil de significativité*

l'autocorrélation de la vitesse ne fonctionne pas car cette autocorrélation reflète principalement des facteurs externes n'apportant pas d'information sur le comportement. Par conséquent, une bonne performance [14] du modèle autoregressif passe par le souci de filtre a priori l'influence des facteurs externes (les courants de surface). Le modèle plus simple de Vermard *et al.* (2010) est plus robuste à ce facteur externe "parasite" [15] que le modèle général.
Parmi les approches HMMs existantes (à savoir Vermard *et al.*, 2010; Walker and Bez, 2010; Peel and Good, 2011; Joo *et al.*, 2013a) il semble donc que celle proposée n'est pas la meilleure, au sens de sa performance de segmentation. Cependant, il est possible que ce fait soit dû au cas d'étude particulier présenté ici. Nous pensons qu'appliquer ce modèle à d'autres données pourrait apporter de nouveaux points de vue sur certaines questions, et mettre en lumière des facteurs intéressants pour l'étude du mouvement, comme les courants de surface ici.

2.4.3 Quel modèle sur la séquence des comportements ?

L'hypothèse Markovienne de transition

Les modèles présentés ici font l'hypothèse que les comportements cachés satisfont la propriété de Markov (équation (2.6)). Pour une chaîne de Markov, cette propriété implique que les temps (discret) de séjour de l'individu dans chacun des comportements suivent une loi géométrique (Norris, 1998a). Il a été montré que cette hypothèse était discutable dans un cadre binaire "pêche/non pêche" (figure 2.16, Bez *et al.*, *in prep*). À la vue de cette figure, une alternative serait de modéliser autrement la durée du temps de séjour dans les états. Cette possibilité est présente dans un modèle de semi-Markov caché (modèle HSMM, Guédon, 2003). Dans ce cas,

14. Pour le problème posé
15. *Idem*

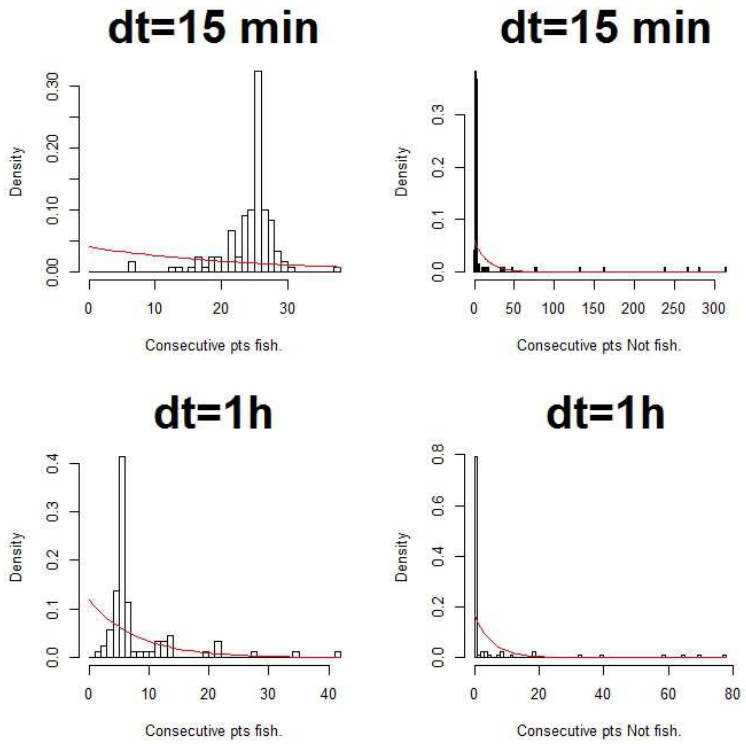


FIGURE *2.16 – Exemples de distributions des temps de séjour dans une activité pour un navire de pêche en Manche. Pour l'exercice, la donnée originale (un pas de temps toutes les 15 minutes), est dégradée au pas de temps 1h. À gauche, la distribution dans l'activité pêche, à droite, dans l'activité non pêche. En rouge, la densité de la loi exponentielle dont la moyenne est égale à la moyenne empirique. Si les états suivaient une chaîne de Markov, la distribution des temps de séjour (en temps continu) serait exponentielle.*

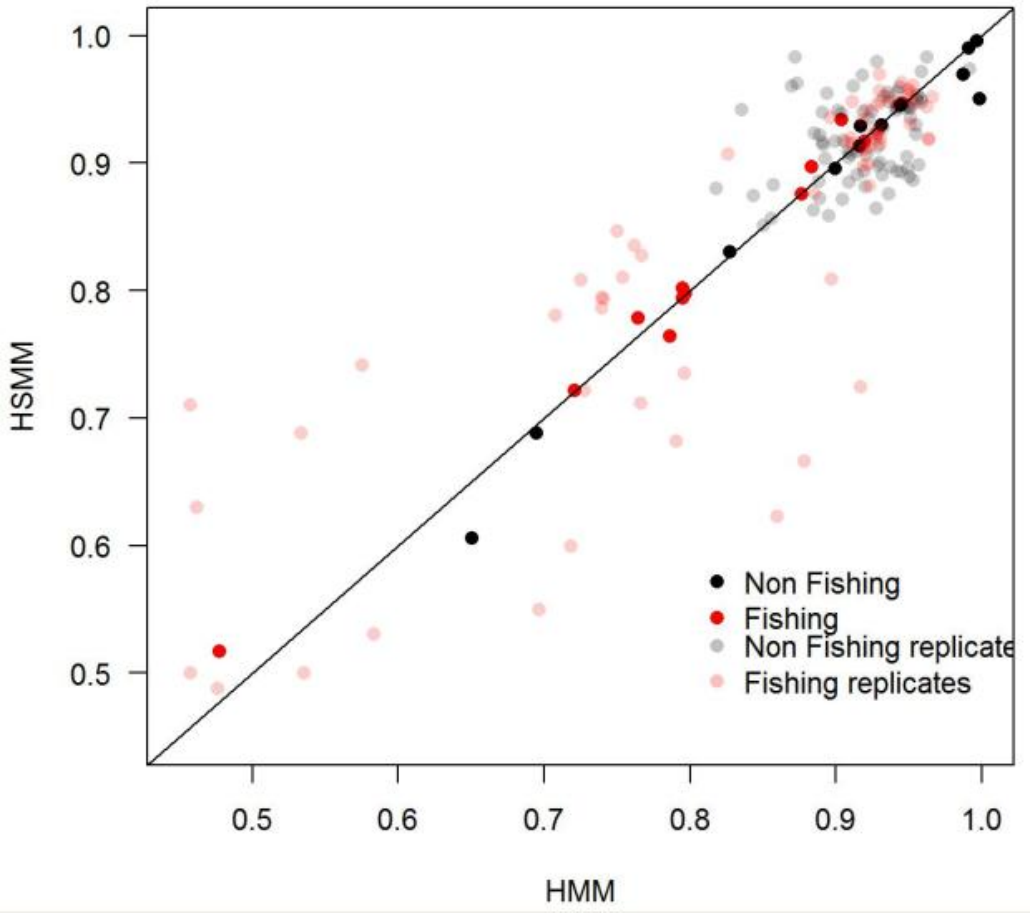


FIGURE *2.17 – Comparaison de la qualité de classification des modèles HMM avec les modèles HSMM. Les observations de positions de différents navires sont classées en pêche/non pêches par des Modèles Markoviens. Les estimations sont comparées avec les réalité obtenues par des observateurs. Les abscisses et les ordonnées donnent le pourcentage de bonne classification par activité. Les points transparent sont issus de réplications par bootstrap (figure issue du travail de Bez* et al.*, in prep).*

ce n'est plus simplement la transition entre deux états qui est modélisée, mais on décrit

$$P(S_{m+1} = i | S_m = j, \ldots, S_{m-d+1} = j) = a_j(d, t_m) p_{ji}(t_m) \tag{2.18}$$

où $a_j(d, t_m)$ est la probabilité de rester d pas de temps dans l'état j au temps t_m considéré, et $p_{ji}(t_m)$ est la probabilité de passer de l'état j à l'état i au temps t_m. On peut alors modéliser la loi de la durée de temps de séjour par la densité de n'importe quelle loi discrète, non nécessairement géométrique. Dans un cadre d'inférence supervisée, pour une classification en 3 comportements sur la pêcherie d'anchois du Pérou, les modèles de Semi-Markov montrent de meilleures performances de segmentation (Joo *et al.*, 2013a). Cependant, pour le cas d'étude présenté ici, les résultats sont similaires à ceux des HMM (voir figure 2.17). Il semble qu'apporter une couche supplémentaire de complexité ne soit pas forcément souhaitable dans le but de détection binaire.

Homogénéité de la loi de transition des comportements

Dans le modèle présenté ici, on travaille avec un chaîne de Markov homogène en temps. Cette hypothèse facilite grandement la problématique de l'inférence. Cependant, cette hypothèse ne semble pas réaliste. En effet, une marée d'un navire de pêche a nécessairement une durée limitée. Ainsi, intuitivement, la probabilité de passer du comportement "pêche" au comportement "route" devrait tendre vers 1 avec le temps. Il serait intéressant de comparer les résultats d'un modèle inhomogène en temps avec les approches présentées ici. Dans le cas où cette hypothèse n'améliorerait pas la segmentation, elle permettrait sans doute de formuler des mécanismes du comportement bien plus réalistes que ceux formulés ici.

Combien de comportements ?

Dans un autre but, ces modèles ont pu être utilisés pour détecter l'utilisation de plusieurs engins de pêche. En effet, dans les pêcheries françaises, il n'est pas rare que des navires utilisent plusieurs engins au cours d'une même marée. Ces engins ne ciblent pas nécessairement les mêmes espèces, et n'ont pas le même effet sur l'environnement. Ainsi, les modèles HMM ont pu être utilisés pour distinguer les opérations de chalutage avec la pêche la ligne, sur des navires français (Woillez *et al*, dans Marchal *et al.*, 2014, Section 1.1.2). Un exemple est montré sur la figure 2.18. Pour tester un modèle à trois états ("Route", "Pêche avec l'engin 1", "Pêche avec l'engin 2"), nous pensons que travailler avec la vitesse par rapport à l'eau est plus approprié. En effet, en augmentant le nombre de comportements, l'impact de facteurs parasites peut être plus grand. Par exemple, la figure 2.19 montre l'ajustement du modèle AR HMM présenté ci dessus avec 3 états, sur un chalutier exerçant en Manche Est. L'interprétation de cet ajustement est complètement différent selon qu'on considère la vitesse par rapport à l'eau, ou par rapport au sol.
Ainsi la question du nombres de comportements à inclure dans le modèle est dépendante du cas d'étude (en halieutique, elle sera dépendante de la pêcherie considérée). Pour comparer deux modèles ayant un nombre de comportements différents, il existe des critères de comparaison afin de sélectionner le meilleur modèle (Jonsen *et al.*, 2013b). Ces critères sont en général des critère de vraisemblance pénalisées.

2.4.4 Une formalisation en temps discret

Le modèle présenté ici (ainsi tous les modèles HMM pour la segmentation en halieutique) est formulé en temps discret. Ces modèles à temps discret requièrent une régularité dans le pas de temps. Plus spécifiquement, le paramètre d'autocorrélation posé dans le modèle AR HMM n'a pas de sens si les données ne sont pas espacées régulièrement. Si pour les études présentées ici, cela n'a pas posé de problème (grâce à la qualité du jeu de données RECOPESCA), il est clair que ce n'est pas adapté à la majorité des cas, où l'irrégularité d'échantillonnage est présente. Dans la pratique, l'irrégularité des données est souvent compensée par une interpolation linéaire (Vermard *et al.*, 2010). Cette interpolation rajoute une deuxième couche d'approximation, après

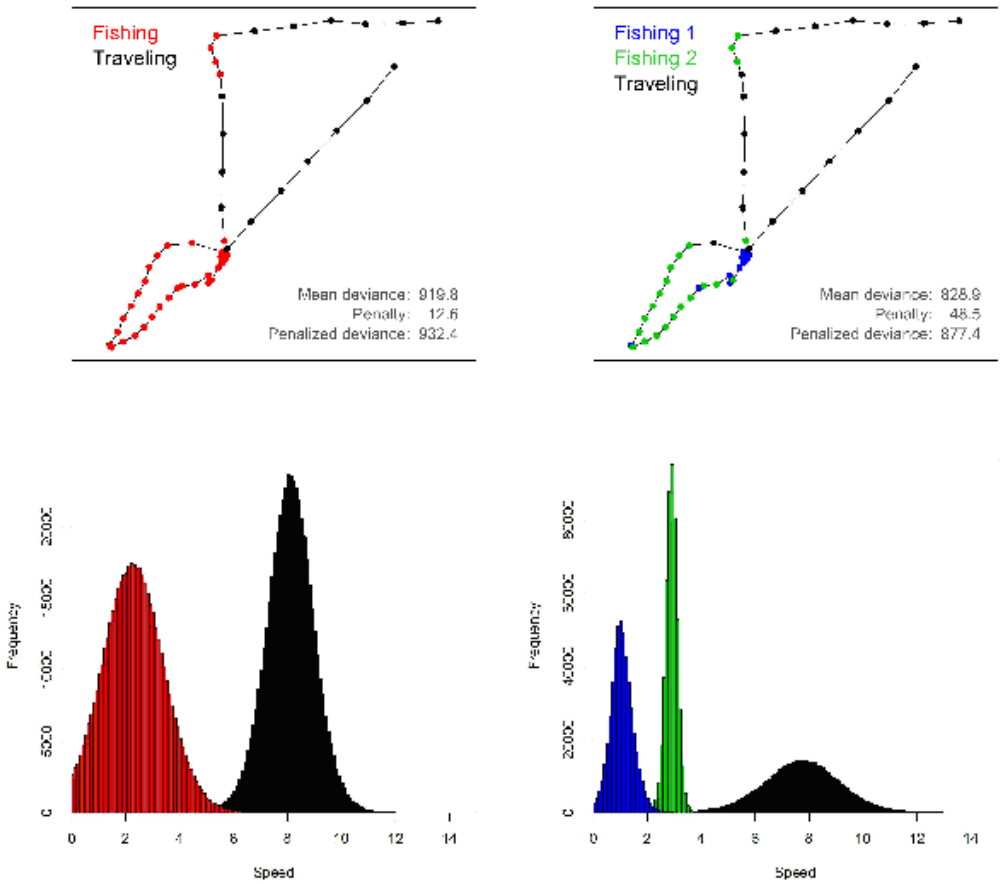


FIGURE *2.18 – Ajustement, dans un cadre Bayésien, d'un modèle HMM à 3 états sur une trajectoire d'un navire utilisant 2 engins, la ligne (correspondant ici à l'enfin de pêche 1. Le modèle HMM utilisé est celui de Vermard* et al. *(2010). À gauche, l'ajustement du modèle avec deux états. À droite l'ajustement du modèle avec 3 états. En haut, les activités estimées du navire au cours de sa trajectoire. En bas, les distributions a posteriori des moyennes des vitesses dans chaque état. Dans l'ajustement à 3 états, on arrive à distinguer l'activité de chalutage, de l'activité de lignage. Figure issue de Woillez et al, dans (Marchal* et al.*, 2014, Section 1.1.2)*

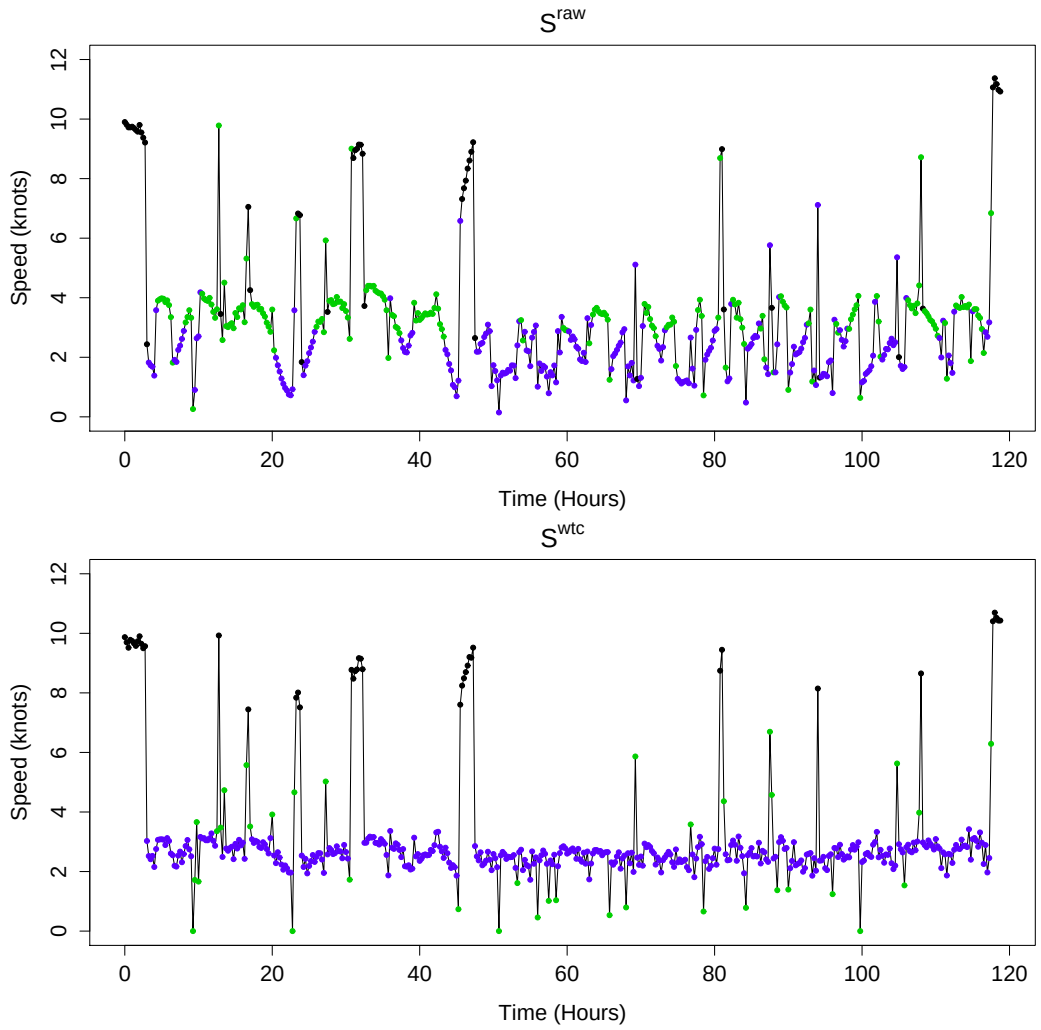


FIGURE *2.19 – Ajustement du modèle AR HMM à 3 états sur une trajectoire. On utilise dans un cas la vitesse par rapport au sol (en haut), dans l'autre cas, la vitesse par rapport à l'eau (en bas). L'interprétation des états est complètement différente. Dans le premier cas, on détecte trois comportements différentiables par leur vitesse. On pourrait interpréter les états comme (i)"Route"(en noir, vitesse rapide) ; (ii)"Pêche avec un engin 1"(en bleu, vitesse lente) ; (iii) "Pêche avec un engin 2" (en vert, vitesse intermédiaire). Dans le deuxième cas, on interprète les comportements comme (i)"Route"(en noir, vitesse rapide) ; (ii)"Pêche"(en bleu, vitesse lente constante) ; (iii) "État transitoire" (en vert, vitesse variable entre le lent, les arrêts, et rapide, les accélérations).*

celle de l'interpolation linéaire pour estimer les vitesses. De même, la formulation des lois de transition du HMM requiert un pas de temps régulier. En effet, les paramètres de la matrice de transition Markovienne tels que décrits ici ne font pas de sens quand le pas de temps est irrégulier.
De plus, la formulation en temps discret suppose que les processus décrits sur la figure 2.2 sont simultanés. Ainsi, dans un tel modèle, les changements de comportement arrivent au moment des observations, ce qui est grandement irréaliste. Enfin, on suppose qu'aucun changement de comportement n'arrive entre deux observations. Cette hypothèse peut être très contraignante quand le pas de temps d'acquisition est long.

2.4.5 Bilan, les HMM, des modèles imparfaits, mais utiles

Nous avons donc vu que les HMM pour la détection d'activité nécessitent de faire des hypothèses sur les mécanismes de la trajectoire. Pour décrire ces mécanismes, des hypothèses statistiques doivent être faites. Ces hypothèses, parfois discutables, comme montré plus haut, permettent de simplifier le problème d'inférence et de se placer dans un cadre statistique connu et maitrisé. D'un point de vue pratique pour l'utilisateur, ce cadre est certainement plus difficile à assimiler que celui adopté dans le choix simple de seuils de vitesses, mais il est très documenté, et maintenant largement implémenté dans les logiciels usuels, comme R (R Core Team, 2014). Même si, à l'heure actuelle de leur développement, les HMMs ne semblent pas faire mieux qu'un simple seuil de vitesse pour détecter l'activité de pêche, il nous semble que ceux-ci ont l'avantage de susciter des questions sur la nature des mécanismes sous jacents à la trajectoire, et à la manière de les formaliser. En ce sens, pour reprendre les mots de McCullagh and Nelder (1989), ils nous semblent "utiles", même s'ils ne doivent pas occulter les autres alternatives.

Chapitre 3

Reconstruire un potentiel de l'environnement à partir de la trajectoire : Une approche par équations différentielles stochastiques

3.1 La trajectoire un objet spatial

Ce chapitre se focalise sur une question différente du chapitre précédent. Nous laissons ici de côté la question de la détection des comportements d'un individu à partir de l'observation de ses trajectoires, pour nous intéresser au lien entre un individu et son environnement.
La manière dont l'environnement influe sur l'utilisation de l'espace d'un individu est une notion clé en écologie, tant dans la volonté de compréhension de la biologie des individus, que dans l'optique d'adopter des mesures de conservation adaptées (Boyce and McDonald, 1999). De ce point de vue, l'analyse des trajectoires des individus peut permettre de mesurer directement l'influence de variables environnementales sur le mouvement adopté. La mesure des variables environnementales permet l'estimation de champs spatiaux qu'il convient de mettre en relation avec les observations des trajectoires d'un individu. Il est alors nécessaire d'extraire un champ spatial des trajectoires. Nous distinguons deux champs spatiaux qu'on peut obtenir à partir de données de trajectoires

- Un champ d'utilisation de l'espace. Ce champ décrit comment l'espace est occupé par l'individu, il résulte de son déplacement ;
- Un champ de perception de l'espace. Ce champ décrit la manière dont l'individu perçoit son environnement. On peut alors voir le déplacement de l'individu comme une conséquence de cette perception.

Le premier type de champ est classique en analyse de trajectoires, et renvoie à la notion de domaine vital ("home range", Burt, 1943), formalisé en termes de densité d'utilisation ("Utilization Distribution" Worton, 1989). Le deuxième champ renvoie à l'idée que le domaine vital est la conséquence d'une carte cognitive de l'individu (Powell, 2000). L'explicitation de ce champ,

et de techniques pour l'estimer, rejoint la volonté des décrire les causes du déplacement, et non d'en décrire seulement les conséquences.
Le reconstruction de ces champs à partir des observations nécessite de spécifier des hypothèses sous-jacentes sur ces dernières. Ces hypothèses permettent la mise en œuvre de techniques d'estimation adaptées, dans un cadre paramétrique ou non.

3.1.1 Quantifier l'utilisation de l'espace

Le déplacement d'un individu résulte en une occupation de l'espace. Un écologue peut alors s'intéresser au temps qu'un individu passe dans une zone en particulier. Cette question est particulièrement utile pour établir des préférences d'habitat par exemple. Dans ces cas, le lien entre le temps passé dans une zone et les caractéristiques de celle-ci sont souvent estimés dans une seconde analyse.

Utilisation globale
Soit un individu Λ se déplaçant dans un espace à deux dimensions entre les temps 0 et T, résultant en une trajectoire $(\mathcal{X}_t)_{0\leq t\leq T}$ [1]. Pour une zone $\mathcal{H}$ de $\mathbb{R}^2$, on peut quantifier la proportion du temps passé par l'individu dans cette zone au cours de l'intervalle $[0,T]$, par la variable aléatoire

$$\frac{1}{T}\int_0^T 1_{\mathcal{H}}(\mathcal{X}_t)dt \tag{3.1}$$

où $1_{\mathcal{H}}(x)$ prend la valeur 1 si $x \in \mathcal{H}$, la valeur 0 sinon. Pour quantifier l'utilisation moyenne de l'environnement, on cherche à estimer l'espérance [2] de cette variable aléatoire. On a alors :

$$\mathbb{E}(\frac{1}{T}\int_0^T 1_{\mathcal{H}}(\mathcal{X}_t)dt) = \frac{1}{T}\int_0^T \mathbb{P}(\mathcal{X}_t \in \mathcal{H})dt \tag{3.2}$$

En notant $p(z,t)$ la densité de probabilité de la variable aléatoire $\mathcal{X}_t$ (variable aléatoire à valeurs dans $\mathbb{R}^2$), on a donc

$$\begin{aligned}\mathbb{E}(\frac{1}{T}\int_0^T 1_{\mathcal{H}}(\mathcal{X}_t)dt) &= \frac{1}{T}\int_0^T \int_{\mathcal{H}} p(z,t)dzdt \\ &= \int_{\mathcal{H}} \left(\frac{1}{T}\int_0^T p(z,t)dt\right)dz. \qquad (3.3)\\ &:= \int_{\mathcal{H}} \tilde{h}(z,T)dz. \qquad (3.4)\end{aligned}$$

La fonction $\tilde{h}(z,T)$ définie par l'équation 3.4 est donc la moyenne sur l'intervalle de temps $[0,T]$ des densités des variables aléatoires $\mathcal{X}_t$, évaluées au point z. Pour tout T, il s'agit une fonction de densité sur $\mathbb{R}^2$ [3]. La fonction $\tilde{h}$ est la densité d'utilisation (Utilization Distribution)

1. En tant que réalisation d'un processus stochastique sous-jacent
2. Par rapport à la loi du processus $\mathcal{X}$
3. En effet, en posant $\mathcal{H} = \mathbb{R}^2$, la variable aléatoire définie en (3.1) vaut 1 pour toute trajectoire. L'espérance définie en (3.2) vaut donc 1 également.

de l'individu (Worton, 1989, dans le cas où h ne dépend pas de T). Le but est de l'estimer à partir des observations des positions d'un individu.

Utilisation conditionnelle (ou locale)

Une autre définition trouvée dans la littérature de la densité d'utilisation (Horne *et al.*, 2007), ne part pas de la variable aléatoire (3.1), mais de la variable aléatoire

$$\frac{1}{T}\int_0^T 1_{\mathcal{H}}(\mathcal{X}_t|X_0^n)dt \tag{3.5}$$

où $X_0^n := (X_0, \ldots, X_n)$ sont les positions acquises aux temps $t_0 = 0, \ldots, t_n = T$. On ne s'intéresse plus à l'utilisation globale de l'environnement par un individu, mais à l'utilisation locale, conditionnellement aux observations de sa position. Autrement dit, sachant que l'individu a été vu en $X_0, \ldots, X_n$, où a-t-il passé son temps entre 0 et T ?

Par le même raisonnement, on peut alors écrire

$$\begin{aligned}\mathbb{E}(\frac{1}{T}\int_0^T 1_{\mathcal{H}}(\mathcal{X}_t|X_0^n)dt) &= \int_{\mathcal{H}}\left(\frac{1}{T}\int_0^T p(z,t,X_0^n)dt\right)dz. \\ &:= \int_{\mathcal{H}} h(z,T)dz. \end{aligned} \tag{3.6}$$

Dans Horne *et al.* (2007), h est également appelée la densité d'utilisation. Cependant, pour faire la distinction avec $\tilde{h}$, nous l'appellerons densité d'utilisation conditionnelle. Encore une fois, il s'agit d'une densité. Le point essentiel est qu'elle est définie conditionnellement aux données.

Deux méthodes sont aujourd'hui principalement utilisées en écologie pour estimer $\tilde{h}$ ou h à partir des trajectoires d'un individu (avec parfois, nous semble-t-il, une confusion entre les deux), la méthode d'estimation par noyau, pour estimer $\tilde{h}$ et la méthode des ponts Browniens, pour estimer h.

Méthode d'estimation par noyau

Dans un article fondateur, Worton (1989) propose l'utilisation d'estimateurs non paramétrique pour estimer la densité d'utilisation $\tilde{h}$ d'un individu à partir d'observations de trajectoires.

On fait alors l'hypothèse que la fonction h définie en (3.6) ne dépend pas de l'intervalle de temps d'observation $[0, T]$ [4]. Cette hypothèse est équivalente à celle d'une densité de probabilité $p(\cdot, t)$ de la variable aléatoire $\mathcal{X}_t$ (présente dans l'équation (3.3)) indépendante de t, et donc que $\tilde{h}(z) = p(z)$ pour tout point $z \in \mathbb{R}^2$ [5].

La suite des observations $(X_0, \ldots, X_n)$ est alors vue comme un vecteur de variables aléatoires identiquement distribuées, de densité p. Cette densité est alors estimée à l'aide de techniques d'estimation non paramétrique de densité.

4. Cette hypothèse, dite de stationnarité, est discutée plus bas (section 3.4.1)
5. La preuve est en annexe C.1

On appelle noyau de densité sur $\mathbb{R}^2$ une fonction $K : \mathbb{R}^2 \mapsto \mathbb{R}$ qui satisfait les conditions suivantes.

Condition de densité $K(x) \geq 0, \int_{\mathbb{R}^2} K(x)dx = 1$;
Condition de symétrie $K(x) = K(-x)$;
Condition de second moment $\int_{\mathbb{R}^2} \| x \|^2 K(x)dx < \infty$;

Un choix courant pour le noyau est la densité d'une Gaussienne, de covariance Σ. On note K_Σ un tel noyau. Pour une trajectoire observée aux points $X_0, \ldots, X_n$ L'estimateur $\hat{h}$ de la fonction $\tilde{h}$ (ou p, ce qui est équivalent ici) est alors

$$\hat{h}_\Sigma(x) = \frac{1}{n+1} \sum_{i=0}^{n} K_\Sigma(x - X_i) \tag{3.7}$$

Problème classique en statistique non paramétrique, le choix de la covariance Σ est critique (et connu sous le nom du problème de choix de fenêtre, ou "bandwith selection problem", Silverman, 1986). Il s'agit du problème principal de ce genre de méthode (Kie *et al.*, 2010). Une méthode classique se base sur la minimisation de la Mean Integrated Squared Error (MISE), on cherche donc $\hat{\Sigma}$ tel que

$$\hat{\Sigma} = \text{argmax}_\Sigma \mathbb{E} \int_{\mathbb{R}^2} (\hat{h}_\Sigma(x) - \tilde{h}(x))^2 dx \tag{3.8}$$

où l'espérance est prise sur la loi des variables aléatoires $X_0, \ldots, X_n$. Cette quantité n'est pas estimable directement, car elle dépend de la fonction $\tilde{h}$, qui est inconnue. Différentes méthodes proposent des approximations pour estimer $\hat{\Sigma}$ à partir de l'équation (3.8) (Seather, 1992).
La détermination de la densité d'utilisation à partir de données de trajectoires n'échappe donc pas au choix de Σ, et le choix de cette fenêtre dépend du problème (et des données) spécifique au cas d'étude.
En analyse de trajectoires, le choix de Σ est impacté par l'autoccorélation des données. En effet, comme dit plus haut, on suppose que les variables aléatoires $X_0, \ldots, X_n$ forment un échantillon identiquement distribué, cependant, en pratique elles ne sont pas indépendantes. Ce problème est d'autant plus vrai que la fréquence des acquisitions GPS est grande (Fleming *et al.*, 2015). Aujourd'hui, la plupart des études font malgré tout l'hypothèse d'indépendance, suivant les premiers travaux de Worton (1989). Récemment, un nouvel estimateur robuste à la haute fréquence a été proposé (Fleming *et al.*, 2015). Cet estimateur fait une forte approximation, en remplaçant $\tilde{h}$, dans (3.8), par la densité d'une loi Gaussienne (Fleming *et al.*, 2015, Appendix B).
Ces méthodes non paramétriques sont aujourd'hui très utilisées car elles sont faciles à mettre en œuvre. À ce titre, elles font donc encore l'objet de recherches visant à les rendre conformes à la réalité des données actuelles. Un exemple d'utilisation des méthodes à noyaux pour estimer la fonction $\tilde{h}$ à partir de données simulées est représenté en figure 3.1. Le noyau utilisé est un noyau Gaussien, la fenêtre Σ est choisie selon l'heuristique proposée par Worton (1989)

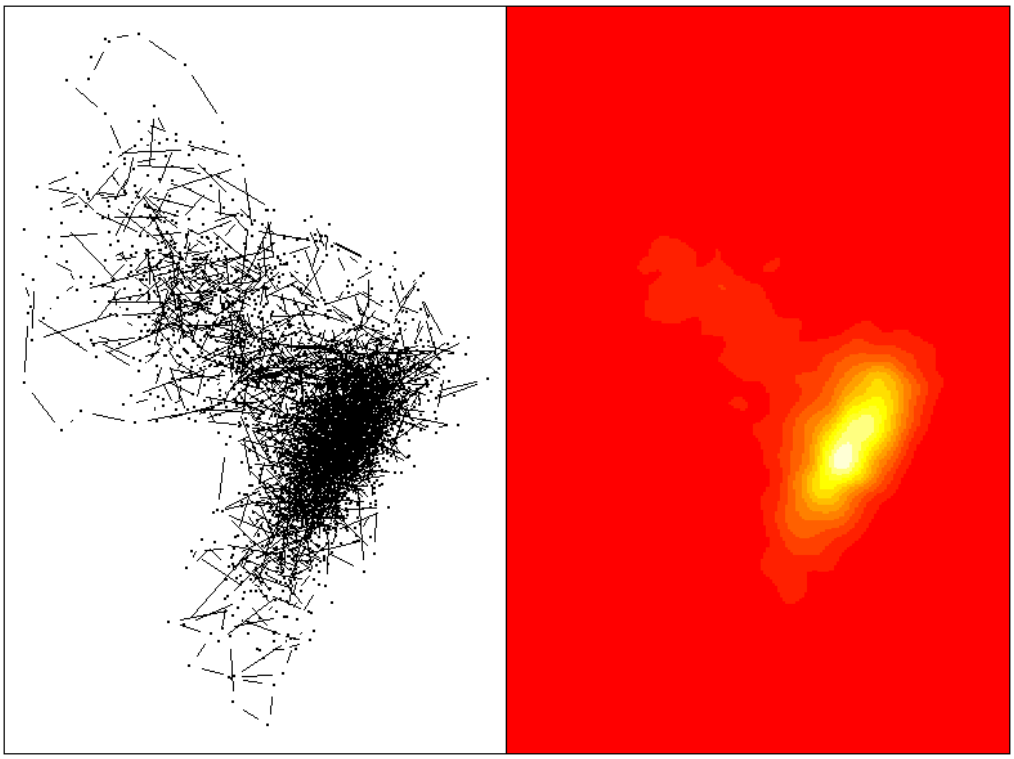


FIGURE *3.1 – Estimation de la densité d'utilisation d'un individu en utilisant la méthode à noyaux proposée par Worton (1989). On a utilisé un noyau Gaussien, en supposant les données indépendantes. La fenêtre Σ est choisie en utilisant une heuristique proposée par Worton (1989). À gauche : les trajectoires observées. À droite, la densité d'utilisation estimée. Les zones rouges représentent des zones faiblement fréquentées, les zones blanches représentent sont fortement fréquentées. L'estimation a été réalisée avec le package R adehabitat (Calenge, 2006).*

Afin de décrire un autre méthode très appliquée en écologie, nous avons besoin de définitions :

Définition 1 (Mouvement Brownien, Mörters and Peres, 2010)**.** *Soit $\mathcal{B} = (\mathcal{B}_t)_{t\geq 0}$ un processus stochastique à valeurs dans $\mathbb{R}^2$ indexé par un temps continu défini sur l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. Alors $\mathcal{B}$ est un mouvement Brownien unitaire si :*

1. *La fonction $t \mapsto \mathcal{B}_t(\omega)$ est continue $\mathbb{P}$-presque sûrement ;*
2. *Les accroissements de $\mathcal{B}$ sont indépendants., i.e., pour $t_1 < t_2$, la variable aléatoire $\mathcal{B}_{t_2} - \mathcal{B}_{t_1}$ est indépendante du processus $(\mathcal{B}_t)_{0\leq t\leq t_1}$.*
3. *Les accroissements de $\mathcal{B}$ sont stationnaires et Gaussiens, i.e., pour $t_1 < t_2$, la variable aléatoire $\mathcal{B}_{t_2} - \mathcal{B}_{t_1}$ est de loi Gaussienne de moyenne nulle et de covariance $(t_2 - t_1)I$. Si la covariance est de la forme $\sigma^2(t_2 - t_1)I$, on parlera d'un mouvement Brownien de paramètre d'échelle σ^2.*

Définition 2 (Pont Brownien)**.** *Soit $\mathcal{B}$ un mouvement Brownien de paramètre d'échelle σ^2 tel que*

$$\mathcal{B}_{t_0} = B_0,\ \mathcal{B}_T = B_T.$$

Le processus $\mathcal{X}^{B_0,B_T,t_0,T} := (\mathcal{B}_t|\mathcal{B}_{t_0} = B_0, \mathcal{B}_T = B_T)_{t_0\leq t\leq T}$. est un pont Brownien de paramètre σ^2. Pour tout t de l'intervalle $[t_0, T]$, la densité de $\mathcal{X}_t^{B_0,B_T,t_0,T}$ est celle d'une loi Gaussienne

telle que :

$$\mathbb{E}(\mathcal{X}_t^{B_0,B_T,t_0,T}) = B_0 + \frac{t-t_0}{T-t_0}(B_T - B_0), \tag{3.9}$$

$$\mathbb{V}(\mathcal{X}_t^{B_0,B_T,t_0,T}) = \sigma^2 \frac{(T-t)(t-t_0)}{T-t_0} I, \ \textit{pour } t_i \leq t_j. \tag{3.10}$$

Le mouvement Brownien est un modèle de mouvement "minimaliste". On suppose un mouvement complètement non dirigé. Ce modèle a été proposé pour décrire le mouvement en biologie par Skellam (1951).

Méthode des ponts Browniens

La méthode des ponts Browniens a été proposée dans sa thèse par Bullard (1999) avant d'être raffinée et popularisée par (Horne *et al.*, 2007). Elle permet d'estimer la densité d'utilisation conditionnelle h définie par l'équation (3.6).

On suppose que la trajectoire de l'individu est la réalisation d'un mouvement Brownien de paramètre d'échelle σ^2. La trajectoire conditionnellement aux observations $X_0^n := X_0, \ldots, X_n$ acquises aux temps $t_0 = 0, \ldots, t_n = T$ est ainsi une suite de réalisations de ponts Browniens $\mathcal{X}^{\mathcal{X}_i,\mathcal{X}_{i+1},t_i,t_{i+1}}$, de paramètre d'échelle σ^2.
Si la trajectoire est observée sans erreur (i.e., $\mathcal{X}_i = X_i$), pour $\mathcal{H} \in \mathbb{R}^2$ on a donc

$$\mathbb{P}(\mathcal{X}_t \in \mathcal{H}|X_0^n) = \mathbb{P}(\mathcal{X}_t \in \mathcal{H}|X_i, X_{i+1}) = \mathbb{P}(\mathcal{X}_t^{X_i,X_{i+1},t_i,t_{i+1}} \in \mathcal{H}), \text{ tel que } t_i \leq t \leq t_{i+1} \tag{3.11}$$

Tel que dit plus haut, cette probabilité est donnée par celle d'une variable aléatoire Gaussienne de paramètres connus (équations (3.9) et (3.10)).
Si la trajectoire est observée avec une erreur, on doit alors poser un modèle d'observation. On modélise la loi de l'observation sachant la vraie position $\mathcal{X}_i$:

$$X_i|\mathcal{X}_i \sim f(\cdot, \mathcal{X}_i, \delta),$$

où densité de probabilité d'une erreur d'observation δ. Dans la pratique, on fait l'hypothèse que l'erreur est distribuée symétriquement autour de $\mathcal{X}_i$ (on fait même souvent l'hypothèse d'une loi Gaussienne). Dans ce cas, on peut alors écrire

$$\mathcal{X}_i|X_i \sim f(\cdot, X_i, \delta).$$

Si on fait l'hypothèse d'indépendance des erreurs, on a alors, pour $t \in [t_i, t_{i+1}]$:

$$\mathbb{P}(\mathcal{X}_t \in \mathcal{H}|X_i, X_{i+1}) = \int\int \mathbb{P}(\mathcal{X}_t^{x,y,t_i,t_{i+1}} \in \mathcal{H}|\mathcal{X}_i = x, \mathcal{X}_{i+1} = y) f(x, X_i, \delta) f(y, X_{i+1}, \delta)\, dxdy, \tag{3.12}$$

où les deux intégrales (par rapport aux variables x et y) sont sur $\mathbb{R}^2$, c'est à dire l'ensemble des possibles pour les vraies positions $\mathcal{X}_i$ et $\mathcal{X}_{i+1}$.

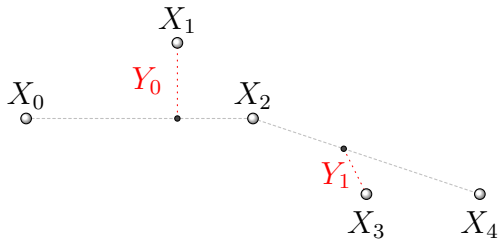


FIGURE *3.2 – Définition des variables Y_i pour la méthode des ponts Browniens (illustration de l'équation* (3.14)*). Le vecteur bivarié Y_1 est l'écart entre l'observation X_1 et son espérance (le point noir sur le segment $[X_0, X_2]$) sous la loi du pont Brownien partant de X_0 et arrivant en X_2. La densité de Y_1 est celle d'une loi normale bivariée de moyenne nulle et de variance donnée par l'équation* (3.15)*, dans le cas où l'observation est sans erreur.*

Que les observations soient faites sans erreur, ou avec, les équation (3.11) et (3.12) permettent de calculer la fonction de densité d'utilisation conditionnelle h. En effet, (en supposant qu'on soit dans le cas avec erreur d'observation), en intégrant chaque terme de l'équation (3.12), pour tous les segments d'observations, on a :

$$\begin{aligned}\int_0^T \frac{\mathbb{P}(\mathcal{X}_t \in \mathcal{H}|X_0^n)}{T} dt &= \int_{\mathcal{H}} \left\{ \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} \iint \frac{p(z, x, y, t_i, t_{i+1}, \sigma^2)}{T} f(x, X_i, \delta) f(y, X_{i+1}, \delta) dx dy dt \right\} dz \\ &:= \int_{\mathcal{H}} h(z, T) dz, \end{aligned} \tag{3.13}$$

où $p(\cdot, x, y, t_i, t_{i+1}, \sigma^2)$ est la densité d'un pont Brownien. Une remarque importante est que les hypothèses faites par la méthode du pont Brownien ne permettent pas de s'affranchir de la dépendance en T de la fonction h [6].
Si σ^2 et δ sont connus, on peut donc calculer la fonction $h(\cdot, T)$ en tout point z de $\mathbb{R}^2$. Cette quantité est en général calculée de manière numérique, nécessitant ainsi le choix d'un pas temporel (en dt) et spatial (en dx et dy) de discrétisation pour calculer l'intégrale entre accolades dans l'équation (3.13). Cependant, des travaux récents ont donné une expression analytique de h quand la trajectoire est observée sans erreur (Van Nieuland *et al.*, 2015).

En général, on suppose δ connu (l'erreur de mesure de l'appareil), le problème d'estimation est donc celui de l'estimation du paramètre σ^2.
Il est estimé en maximisant une pseudo vraisemblance. En considérant observations impaires, i.e. les observations $X_1, X_3, \ldots$. On calcule

$$Y_k = X_{2k+1} - \left(X_{2k} + \frac{t_{2k+1} - t_{2k}}{t_{2(k+1)} - t_{2k}} (X_{2(k+1)} - X_{2k}) \right), \quad k = 0, \ldots, [n/2] \text{ [7]} \tag{3.14}$$

La figure 3.2 montre une interprétation géométrique des variables aléatoires Y_k. Sous les hy-

6. Nous reviendrons sur cet aspect dans la section 3.4.1
7. La partie entière de $n/2$

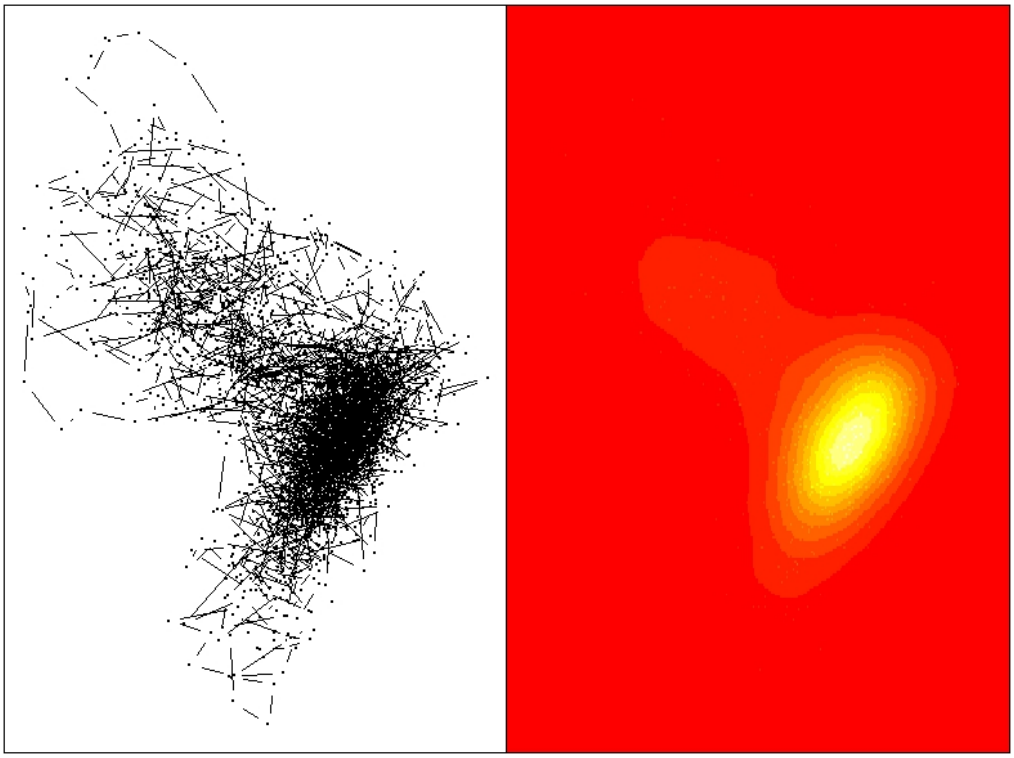


FIGURE *3.3 – Estimation de la densité d'utilisation conditionnelle d'un animal en utilisant la méthode des ponts Browniens. L'estimation a été réalisée avec le package R BBMM (Nielson* et al.*, 2013)*

pothèses du pont Brownien, les variables aléatoires Y_k sont des Gaussiennes indépendantes, de moyenne nulle et de variance (quand l'observation est sans erreur) :

$$\sigma^2 \frac{(t_{2(k+1)} - t_{2k+1})(t_{2k+1} - t_{2k})}{t_{2(k+1)} - t_{2k}} I. \tag{3.15}$$

Une équation similaire à (3.15) existe quand l'erreur d'observation est non nulle (Horne *et al.*, 2007). On peut alors calculer le σ^2 qui maximise la vraisemblance des $(Y_k)_{k=0,\dots,[n/2]}$. Différents ajouts ont été faits à la méthode des ponts Browniens. L'utilisation d'un pont Brownien avec dérive a été proposée pour tenir compte du choix de direction de l'individu (Benhamou, 2011), une méthode de détection de rupture a été utilisée pour détecter des changements dans le σ^2 au cours de la trajectoire (Kranstauber *et al.*, 2012), et une extension pour des marche aléatoires anisotropes (i.e. des paramètres d'échelle σ_x^2 et σ_y^2 pour les deux directions de l'espace) a été proposée (Kranstauber *et al.*, 2014). Un exemple jouet d'application des ponts Browniens pour estimer la densité d'utilisation conditionnelle, avec la méthode proposée par Horne *et al.* (2007), est montré sur la figure 3.3.

3.1.2 Lien entre utilisation et mouvement

Nous avons vu deux méthodes pour quantifier l'utilisation de l'espace. Dans la méthode d'estimation non paramétrique de cette densité d'utilisation par noyaux, aucun modèle paramétrique n'est supposé pour le mouvement. Cependant, une hypothèse très forte de stationnarité est faite pour l'individu. Cette hypothèse implique que la densité de la probabilité de présence

p d'un individu ne dépend pas du temps. Ainsi, on suppose que les positions d'un individu sont identiquement distribuées (mais potentiellement corrélées) d'une même loi p. Intuitivement, cela implique que l'utilisation de l'espace par l'individu est dans un état d'équilibre. Cependant, aucun mécanisme n'est explicité pour expliquer l'atteinte de cet équilibre.
La méthode des ponts Browniens ne fait pas l'hypothèse d'un état d'équilibre dans l'utilisation de l'espace. On s'intéresse donc seulement à l'utilisation de l'espace faite par l'animal entre les moments d'observation. Ce modèle fait une hypothèse forte sur le mouvement, en le supposant complètement indépendant de la manière dont il perçoit l'environnement. On n'a donc aucune idée de ce qui guide le déplacement. De plus, les propriétés du mouvement Brownien font que cette densité d'utilisation conditionnelle calculée n'a qu'un caractère local, et ne peut apporter aucun caractère prédictif sur les positions de l'animal au cours du temps.

3.1.3 Quantifier la perception de l'espace

Une autre vision sur le lien entre le mouvement et l'espace est la vision mécaniste des domaines vitaux :

> *« [Home range] models are mechanistic in the sense that the pattern of space use by an animal is calculated by an explicit mathematical scaling of [...] underlying rules of movement. Thus, unlike statistical home range models, the patterns of space used obtained from mechanistic home range models are not arbitrary distributions but rather reflect the pattern of space use that results from the underlying set of rules governing the individual's movement. »*
>
> **Moorcroft and Lewis (2013)**

L'approche mécaniste de l'étude des domaines vitaux (ou des densité d'utilisation) se doit d'expliciter les mécanismes guidant l'utilisation de l'habitat. Contrairement à l'estimation directe de la densité d'utilisation (globale ou conditionnelle), l'explication des mécanismes permet ainsi de reproduire un mouvement, au travers d'un modèle. Cette approche a donc un potentiel prédictif, si les mécanismes du mouvement sont bien retranscrits.

De ce point de vue, une modélisation possible est de construire un champ spatial qui est une cause directe du mouvement et qui détermine ainsi l'émergence de la densité d'utilisation telle que définie plus haut. L'objectif est donc de définir le mouvement à partir de ce champ spatial, et d'estimer ce dernier à partir des observations du déplacement.

> *« La cause du mouvement est en définitive le désir qui est produit, soit par la sensation, soit par l'imagination, soit par l'intelligence. »*
>
> **Aristote, Le mouvement des animaux**

Nous pensons ici que ce qui guide, en moyenne, le mouvement d'un individu est son intérêt. Ainsi, un individu va se déplacer en moyenne vers les zones qu'il préfère. La préférence de ces zones peut avoir des raisons multiples (la "sensation", "l'imagination", "l'intelligence"). Ainsi, nous pensons qu'en moyenne, un individu va aller vers des zones attractives pour lui. Dans la

section suivante, on se propose de formaliser ce concept au travers des équations différentielles stochastiques (EDS).

Un formalisme d'EDS pour le modèle de déplacement

Ici, nous nous appuyons sur le formalisme du modèle de déplacement introduit par Brillinger *et al.* (2001a,b); Brillinger (2010); Preisler *et al.* (2004, 2013), et déjà mis en oeuvre pour modéliser le déplacement de mammifères marins et terrestres.
Nous partons du principe qu'un individu attribue à chaque point de l'espace une valeur de potentiel attractif. Pour un individu, la manière dont il perçoit son environnement au temps t est donc modélisée par une fonction de potentiel :

$$\begin{aligned} P : \mathbb{R}^2 &\mapsto \mathbb{R} \\ x &\mapsto P(x,t) \end{aligned}$$

Cette fonction peut être à valeurs positives ou négatives. Les zones attractives seront les zones des plus fortes valeurs prises par P.
La deuxième hypothèse est que l'individu, en moyenne, a tendance à se déplacer vers les zones attractives. Ainsi, en moyenne, la direction du déplacement d'un individu, situé en x au temps t, va être le gradient de la surface P, évaluée en x et t. Supposons que la trajectoire $\mathcal{X}$ d'un individu commence en $\mathcal{X}_0$. En mécanique déterministe, la dérivée de la position $\mathcal{X}_t$ serait donc donnée par

$$\frac{d\mathcal{X}_t}{dt} = \nabla P(\mathcal{X}_t, t) \tag{3.16}$$

Ce qui aboutirait, pour connaître la position, à l'équation

$$\mathcal{X}_t = \mathcal{X}_0 + \int_0^t \nabla P(\mathcal{X}_s, s) ds \tag{3.17}$$

Or, pour des raisons évoquées en introduction (section 1.3), la modélisation se fait ici dans un cadre stochastique.
Pour modéliser continûment au travers du temps un phénomène stochastique, on utilise le formalisme des équations différentielles stochastiques. Ce formalisme demande de nombreux outils mathématiques que nous n'expliciterons pas ici. Cependant, nous reprenons ici une partie de l'introduction "appliquée" aux EDS proposée par Särkkä (2012). Nous pensons que cette introduction peut aider à comprendre l'intuition derrière les EDS [8].
Dans un cadre stochastique, le déplacement infinitésimal n'est pas déterministe, l'idée est alors de réécrire (3.16) de la manière suivante

$$\frac{d\mathcal{X}_t}{dt} = \nabla P(\mathcal{X}_t, t) + \Gamma(\mathcal{X}_t, t) w_t \tag{3.18}$$

8. Et toute la difficulté pour une formulation complète et rigoureuse !

où w_t est un bruit blanc, et $\Gamma(\mathcal{X}_t, t)$ est une matrice de covariance. L'équation (3.18) introduit une stochasticité dans le déplacement au niveau infinitésimal. L'intégration sur le temps de l'équation (3.18) peut se faire de manière formelle, et elle définit l' équation différentielle stochastique (EDS)

$$\mathcal{X}_t = \mathcal{X}_0 + \int_0^t \nabla P(\mathcal{X}_s, s)ds + \int_0^t \Gamma(\mathcal{X}_s, s)dW_s \tag{3.19}$$

ou, son équivalent avec la notation différentielle

$$\mathcal{X}_0 = \mathcal{X}_0,\ d\mathcal{X}_t = \nabla P(\mathcal{X}_t, t)dt + \Gamma(\mathcal{X}_t, t)dW_t \tag{3.20}$$

Dans l'équation (3.19), le dernier terme du membre de droite est l'intégrale d'Itô. Elle est construite formellement à l'aide du mouvement Brownien, et ne sera pas définie ici [9]. Dans les équations (3.19) et (3.20), W est un mouvement Brownien [10], moralement, le bruit blanc infinitésimal w_t de (3.16) peut être considéré comme la "dérivée formelle" au temps t du mouvement Brownien [11].
La fonction ∇P est la fonction de dérive de l'EDS, la fonction Γ est la fonction de diffusion. Le modèle de déplacement est ainsi complètement déterminé par l'équation (3.20).
La solution de l'équation (3.20), si elle existe, est un processus stochastique. On considère donc que la trajectoire d'un individu est une réalisation d'un tel processus. Les trajectoires représentées sur les figures 3.1 et 3.3 ont été simulées selon la loi d'un tel processus. La fonction P correspondante est représentée sur la figure 3.4.

Forme de la fonction P

Pour garantir l'existence d'une solution unique à l'EDS, les fonctions de dérive et de diffusion doivent satisfaire certaines conditions [12]. L'équation du mouvement impose que P soit une fonction dérivable. Cette hypothèse est forte, elle signifie qu'un individu perçoit son environnement de manière lisse. Il n'existe pas de changements abrupts dans la manière dont l'environnement est perçu.
Pour un point G de l'espace, si la fonction P est de la forme $P(x,t) \propto \| x - G \|^2$, et $\Gamma(x,t) = \sigma I$, alors le processus solution de (3.20) est le processus de Ornstein Ulhenbeck, ou processus autoregressif continu (Uhlenbeck and Ornstein, 1930). Le point G est alors un point attractif, que

9. On renvoie le lecteur à Øksendal, 2003 pour une définition rigoureuse de l'intégrale d'Itô, et des équations différentielles stochastiques

10. Ou processus de Wiener, d'où le W

11. Cette intuition est celle de l'auteur, est n'est pas rigoureuse, encore une fois, voir Øksendal (2003)

12. En posant $b(x,t) = \nabla P(x,t)$, les deux conditions suivantes sont suffisantes pour l'existence d'une unique solution continue (Øksendal, 2003) :

Lipschitz Il existe $K < \infty$ tel que pour tout $x, y, \in \mathbb{R}^2$ et $t \in [0,T]$:

$$\| b(x,t) - b(y,t) \| + \| \Gamma(x,t) - \Gamma(y,t) \| < K \| x - y \|$$

Croissance linéaire Il existe $C < \infty$ tel que pour tout $x, y, \in \mathbb{R}^2$ et $t \in [0,T]$:

$$\| b(x,t) \| + \| \Gamma(x,t) \| < C(1 + \| x \|)$$

l'on tend à rejoindre depuis n'importe quel point de l'espace. Ce processus présente l'avantage substantiel d'être un processus Gaussien [13], dont la loi est connue. La carte de perception de l'espace est alors une fonction unimodale. Ce modèle et des extensions ont été proposés pour décrire le mouvement d'individus dans un environnement composé de multiples habitats (Blackwell, 1997; Harris and Blackwell, 2013).
Dans l'optique où P est la valeur qu'attribue un individu à son environnement, se pose la question sur les valeurs prises par P. On peut naturellement penser qu'un individu n'attribue pas une valeur arbitrairement grande (négative ou positive) à un point de l'espace. Cette condition n'est pas satisfaite quand on étudie un processus de Ornstein Ulhenbeck, dont le potentiel associé est non borné. D'un autre côté, on peut penser qu'une barrière infranchissable ou un environnement invivable peut se voir attribuée une valeur aussi grande (négativement) que voulue.
Dans la suite, nous proposons l'utilisation de fonctions P bornées pour modéliser le potentiel perçu de l'environnement.

Problématique d'inférence

À partir de maintenant, nous nous plaçons dans un cadre d'équation différentielle stochastique homogène, i.e. :

$$P(x,t) = P(x);\ \Gamma(x,t) = \Gamma$$

Dans ce cas, le processus $\mathcal{X}$ solution de (3.20) est un processus Markovien homogène, i.e. :

$$p\left(\mathcal{X}_{t+\Delta}|(\mathcal{X}_s)_{0\leq s\leq t}, [0, t+\Delta]\right) = p(\mathcal{X}_{t+\Delta}|\mathcal{X}_t)$$

En pratique, on va choisir une forme fonctionnelle pour P et Γ, dépendant de paramètres θ et γ qu'il faudra donc estimer. Pour la majorité des fonctions $P(\cdot,\theta)$ [14], pour deux observations $X_t, X_{t+\Delta}$ discrètes du processus solution de (3.20), la log vraisemblance

$$p(\theta, \gamma|X_t, X_{t+\Delta})$$

n'est pas calculable directement, car la loi de transition n'a pas de forme analytique.
Dans leur article, Preisler *et al.* (2013) utilisent un schéma d'Euler pour approcher la vraisemblance. Dans ce cas là, la densité de la loi de transition $X_{t+\Delta}|X_t$ est approchée par la densité d'un loi normale $\mathcal{N}(X_t + \nabla P(X_t)\Delta, \Delta\Sigma(X_t))$. Cette approximation permet de calculer facilement une approximation de la vraisemblance. Cependant, cette approximation n'est valable que quand Δ tend vers 0. L'erreur commise quand le pas de temps est long n'est pas quantifiable, et aucune étude empirique n'a été proposée pour la quantifier en analyse de trajectoires en écologie.
Une autre approche, encore non utilisée en écologie, peut être développée. Elle se base sur le

13. Si $\mathcal{X}_0$ est Gaussien, ou connu
14. À l'exception de quelques cas, dont le mouvement Brownien, et le processus de Ornstein Ulhenbeck

théorème de changement de mesure de Girsanov [15]

Théorème 1 (Girsanov)**.** *Soit l'équation différentielle stochastique*

$$d\mathcal{X}_t = \alpha(\mathcal{X}_t, \theta)dt + dW_t,\ \mathcal{X}_0 \text{ connu},\ 0 \leq t \leq T \tag{3.21}$$

où la forme de $\alpha(\cdot, \theta)$ garantit l'existence d'une solution [16]*. Soit $\mathbb{Q}_\theta^{\mathcal{X}_0,T}$ la mesure de probabilité induite par la solution de cette EDS. Soit $\mathbb{W}^{\mathcal{X}_0,T}$ la mesure de probabilité induite par un mouvement Brownien partant de $\mathcal{X}_0$ entre 0 et T. Alors, ces deux mesures sont équivalentes et on a, pour tout chemin continu $\mathcal{X}$ de $[0,T]$ dans $\mathbb{R}^2$, partant de $\mathcal{X}_0$:*

$$\frac{d\mathbb{Q}_\theta^{\mathcal{X}_0,T}}{d\mathbb{W}^{\mathcal{X}_0,T}}(X) = \exp\left\{\int_0^T \alpha(\mathcal{X}_s, \theta)d\mathcal{X}_s - \frac{1}{2}\int_0^T \parallel \alpha(\mathcal{X}_s) \parallel^2 ds\right\}. \tag{3.22}$$

Supposons qu'on dispose d'une observation continue $\mathcal{X} = (\mathcal{X}_t)_{0\leq t\leq T}$ du processus solution de (3.21). Dans ce cas, pour un θ donné, le calcul du terme de droite [17] de l'équation (3.22) donne la vraisemblance de θ selon cette observation. Cette vraisemblance est exprimée comme une densité par rapport à la mesure induite par mouvement Brownien, ou mesure de Wiener. Cependant, dans la pratique, on ne dispose pas d'une observation continue. On se retrouve donc dans un cadre aux données manquantes, où la donnée manquante serait tout le chemin parcouru entre 0 et T. Comme dans la section 2.1.3, quand on ne dispose que d'observations discrètes de la trajectoire, on peut développer un algorithme de type EM qui permet de maximiser la vraisemblance du modèle [18].

15. La version présentée ici est une version édulcorée de celle présentée dans Iacus (2009).

16. Si $\alpha(,\theta)$ est dérivable est bornée, une unique solution existe.

17. Ce calcul nécessite tout de même le calcul de l'intégrale d'Itô en $d\mathcal{X}_s$, la section suivante donne une condition pour que cette intégrable soit calculable en deux dimensions

18. Une remarquable référence pour les méthodes d'inférence de processus de diffusion est la thèse de Sermaidis (2010). Cette thèse ajoute de très bonnes explications intuitives aux explications formelles des algorithmes d'estimation qu'elle présente

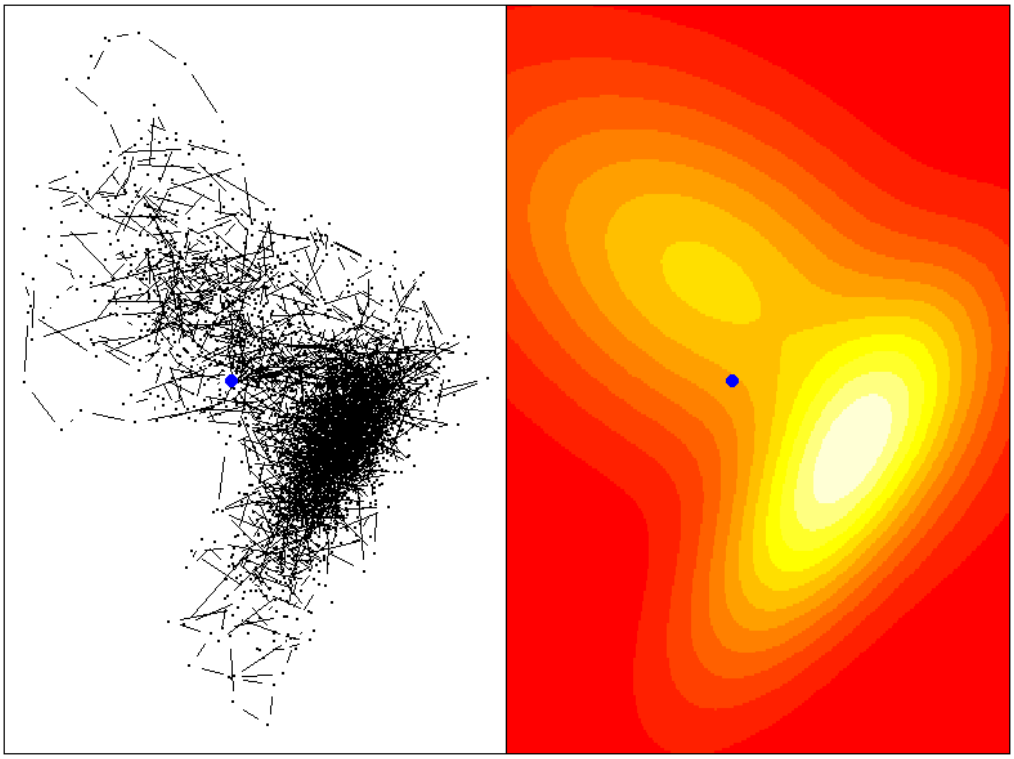


FIGURE *3.4 – Exemple de trajectoires simulées selon la loi du processus solution de* (3.20)*. La fonction P dont le gradient définit la dérive de l'EDS est représenté à droite. Le point bleu représente le points de départ des 30 trajectoires simulées.*

3.2 Un modèle continu basé sur un potentiel de forme Gaussienne : le modèle GaP

Dans la section précédente, nous avons décrit ce que serait un modèle de mouvement où l'individu est guidé par une carte de potentiel. En fin de section, nous avons donné une intuition sur la problématique de l'inférence.
Dans cette section, un tel modèle est proposé pour l'halieutique. Ce modèle est basé sur un potentiel dont la forme est celle d'un mélange de fonctions dites "Gaussienne" (au sens où elles sont de la forme $\exp(-x^2)$). Par la suite, on se réfèrera à ce modèle comme le modèle GaP (Gaussian Potenial).
L'estimation par maximum de vraisemblance se fait par un algorithme de type EM. L'étape E ne pouvant être réalisé analytiquement, elle est approchée par méthode de Monte Carlo. L'approche Monte Carlo se base sur la simulation conditionnelle du processus par les algorithmes exacts développés dans Beskos and Roberts G. (2005); Beskos *et al.* (2006b). La procédure de maximisation de la vraisemblance est ici entièrement détaillée.
Ce travail a donné lieu à une soumission d'article dans la revue Journal of Royal Statistical Society, series C (Applied Statistics). Il a été réalisé avec Marie-Pierre Étienne et Sylvain Le Corff.
L'auteur présente une dernière fois ses excuses aux lecteurs pour le passage à l'anglais.

Stochastic differential equation based on a Gaussian potential field to model fishing vessels trajectories

Pierre Gloaguen* Marie-Pierre Etienne‡ Sylvain Le Corff§

3.2.1 Introduction

In ecology, studies on the movement of animals provide many insights on the ecological features that explain population-level dynamics. These analyses are crucial to wildlife managers to understand complex animal behaviors Chavez (2006). In fisheries science, understanding the underlying patterns responsible for spatial use of the ocean is a key aspect of a sustainable management Chang (2011).

Both fields promote now large programs to deploy Global Positionning System (GPS) device. For instance we might mention, among others, the Tagging of Pelagic Predators program (TOPP) [19] or the TORSOOI [20] program for marine animals and the WOLF GPS [21] or Elephant

19. www.gtopp.org
20. www.torsooi.com
21. www.wolfgps.com

without Borders [22] programs for terrestrial wildlife. In the European Union, since the 1st of January 2012, the fishing vessels above 12m are mandatory equipped with a Vessel Monitoring System which has become a standard tool of fisheries monitoring and control worldwide. As a result such programs produce large amount of trajectory data. These data sets have been largely used to understand, explain and potentially predict animals/vessels movements.

Those tasks require modeling and analysis of the GPS tracks data but may have different objectives. A first objective is to segment the whole trajectory in homogeneous regions which are to be linked with behavioural activities. This is classically addressed using Hidden Markov Models Joo *et al.* (2013b); Gloaguen *et al.* (2015) or change point detection approaches Barraquand and Benhamou (2008). A second objective is the construction/definition of a land/space use map, defined as Utilization Distribution (UD) of the individual, using GPS data. Two main approaches are used to build these UD maps. First, Nonparametric kernel methods usually require assumptions such as independent data that depend strongly on the sampling scheme of the data (although these assumptions have been relaxed to account for autocorrelation in some recent works Fleming *et al.* (2015)). Another approach, based on Brownian Bridge techniques, assumes that the trajectory is a Brownian motion and then builds the UD by integrating the probability of presence over time.

These two approaches rely on unrealistic assumptions regarding the movement dynamics. In this paper, it is assumed that observed trajectories of an individual are direct consequences of the environment attractiveness, hereafter called potential. We develop a method based on continuous time and continuous space stochastic modeling to estimate potential maps. The random process of interest $(X_t)_{t\geq 0}$ describes the position of an individual continuously in time and space but discretely observed. This process is assumed to be a solution to a stochastic differential equation (SDE) which depends on the environment potential. This statistical framework allows to (a) estimate the parameters of the SDE and (b) sample trajectories of individuals to predict future behaviors.

The use of a special form of SDE to model animal movement have been studied in Johnson *et al.* (2008) and Harris and Blackwell (2013). These papers focus on the mean-reverting Ornstein-Uhlenbeck process and its extensions. The Ornstein Ulhenbeck process has a known distribution that facilitates parameters estimation. However, it is a quite restrictive class of model, and its extensions make the estimation framework more complex. Preisler *et al.* (2004) and Preisler *et al.* (2013) consider the special case where the drift of the SDE is the gradient of a potential function which is a weighted sum of different sources that define attractive (resp. repulsive) regions where the individuals are likely (resp. unlikely) to move. As the solution to the corresponding SDE does not have an explicit density, the drift function is estimated by applying an Euler scheme to interpolate trajectories between two consecutive observations.

22. `www.elephantswithoutborders.org/tracking.php`

This introduces an intrinsic bias in the estimation procedure, and might not be well suited for tracking data that can be sparsely sampled.

In this paper, a SDE based on a potential function is proposed to model the position of an individual at each time step t. The drift of the SDE is a mixture of Gaussian shaped functions, with unknown weights, centers and shapes, which represent the attractive regions where the species/fishing vessels are likely to travel.

The aim of this paper is to find the maximum likelihood estimator (MLE) for these attractive regions using GPS data. Although the SDE has no explicit solution, considering G independent animal/vessel trips $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^G$, a trip $\boldsymbol{X}^g$ being a sequence of observations $\boldsymbol{X}^g = (X_0^g, \ldots, X_{n^g}^g)$ sampled at times $t_0^g, \ldots, t_{n^g}^g$, the parameters of the SDE (weights, centers and shapes of the attractive regions) can be estimated using an Expectation Maximization (EM) based algorithm. The E step is approximated by Monte Carlo methods, using the exact algorithms introduced in Beskos *et al.* (2006a) and Beskos and Roberts G. (2005) for an exact sampling of the SDE. The M step is performed using the gradient free CMAES approach described in Hansen and Ostermeier (2001). As this proposed method does not rely on a discrete scheme to approximate the true process, the estimation error of the MLE only depends on the quality of the Monte Carlo approximation of the E step.

The paper is organized as follows. In Section 3.2.2, the target SDE is introduced and the EM procedure based on independent trips to estimate the parameters is displayed in Section 3.2.3. Performance of the proposed algorithm is assessed in Section 3.2.4 with simulated data and in Section 3.2.4 using a real data set. Technical results to apply the EA algorithm of Beskos *et al.* (2006a) and Beskos and Roberts G. (2005) are postponed to Appendix D.1 and D.2.

3.2.2 Model and objectives

The goal of this paper is to propose a model which allows to identify regions of high attractiveness for an individual using GPS tagging. Those regions may then be ecologically interpreted and understood as feeding zones for different animals or high concentration of commercially interesting fishes for fishing vessels. The movement is modeled using a SDE on $\mathbb{R}^2$. The drift is the gradient of a potential map which value at location x represents the attractiveness. The diffusion term is assumed to be a constant scalar matrix. This apparently quite restrictive assumption is motivated by technical reasons but seems also quite reasonable due to a lack of biological information on the stochastic part of the movement.
Formally, the model is defined as follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let C be the set of continuous maps from $[0, T]$ ($T > 0$ is a fixed time horizon) to $\mathbb{R}^2$, called the Wiener space, endowed with the usual σ-algebra $\mathcal{C}$ generated by cylinders. In the following, we consider the unique measure $\mathbb{W}_T$ on the Wiener space such that the coordinate process $(W_t)_{0 \leq t \leq T}$ is a standard Brownian motion and the natural filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Each observed trajectory $\mathbf{X}^g$ is

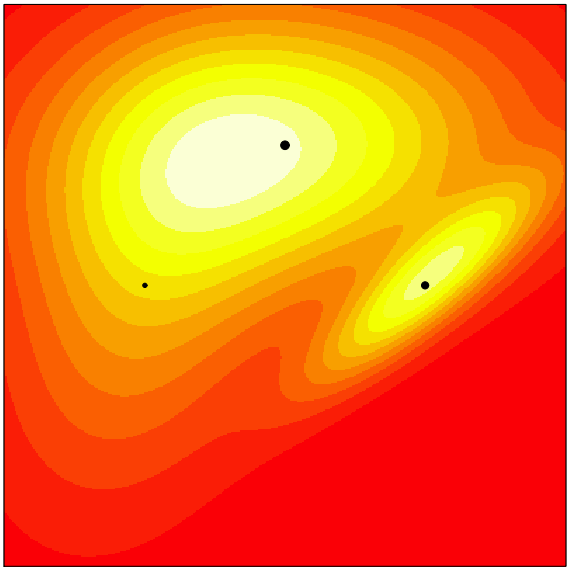


FIGURE *3.5 – One example of potential map (when $K = 3$). Each black dot represents the position of μ_k ($k = 1, 2, 3$), the dot size being relative to the weight π_k.*

a realization of the position process $(X_t)_{t\geq 0}$ which is assumed to be a solution to the following homogeneous SDE :

$$X_0 = X_0^g \quad \text{and} \quad \mathrm{d}X_t = b_\theta(X_t)\mathrm{d}t + \gamma \mathrm{d}W_t\,, \tag{3.23}$$

where $\gamma \in \mathbb{R}$ is the diffusion coefficient and the drift function b_θ is defined as follows : $b_\theta := \nabla P_\theta$, where $P_\theta : \mathbb{R}^2 \to \mathbb{R}$ is given, for all $x \in \mathbb{R}^2$, by

$$\begin{aligned} P_\theta(x) &:= \sum_{k=1}^{K} \pi_k \varphi_k^\theta(x)\,, \\ \varphi_k^\theta(x) &:= \exp\left(-\frac{1}{2}(x-\mu_k)^T C_k (x-\mu)\right). \end{aligned} \tag{3.24}$$

where $\pi_k \in \mathbb{R}, \mu_k \in \mathbb{R}^2$, and C_k is a positive-definite matrix. The parameter θ contains all the unknown quantities we wish to estimate : $(\mu_k)_{1\leq k\leq K}$, $(C_k)_{1\leq k\leq K}$ and $(\pi_k)_{1\leq k\leq K}$ and we will discuss either γ has to be estimated or not. μ_k, C_k and π_k are respectively location, shape and weight parameters of the k-th attractive zone. The number of attractive zones K is first assumed to be known. Therefore, for all $x \in \mathbb{R}^2$,

$$b_\theta(x) = -\sum_{k=1}^{K} \pi_k \varphi_k^\theta(x) C_k\,(x-\mu_k)\,. \tag{3.25}$$

Since b_θ is Lipschitz and γ is constant, the solution to this SDE exists and is unique.
Figure 3.5 shows one possible example of map of space use, the gradient of which defining the drift b of the proposed SDE. The model therefore captures the idea of attractive zones we aim at identifying. The parametric form of (3.24) provides a smooth and flexible framework to describe different potential maps. The potential map is chosen positive (but could be chosen negative), but has no constraint to integrate to 1. For these reasons, the map is not a probability

distribution. However, as the potential is supposed to be bounded, it might be understood as a measure of attractiveness.
The assumption on the diffusion parameter is quite strong. In a more general context we could have defined the diffusion term as a matrix Γ but we would need to ensure that the proposed model is such that the function $x \mapsto \Gamma^{-1} b(\Gamma x)$ is conservative, i.e. such that there exists

$$H_\theta : \mathbb{R}^2 \to \mathbb{R} \text{ satisfying } \nabla H_\theta = \Gamma^{-1} b_\theta(\Gamma x). \tag{3.26}$$

In the model described by (3.23) and (3.25), the mapping $x \mapsto \gamma^{-1} b(\gamma x)$ is conservative with

$$H_\theta : x \mapsto \sum_{k=1}^{K} \pi_k \varphi_k^\theta(\gamma x)/\gamma \,. \tag{3.27}$$

This conservative property is crucial to apply the exact algorithm 1 of Beskos *et al.* (2006a), which allows to sample skeleton of trajectories exactly distributed as the solution to (3.23). This assumption could be modified in the case of a non constant diffusion matrix $x \mapsto \Gamma(x)$ at the cost of a lot of technical complications.

This model depends on the parameters γ, $(\pi_k)_{1\leq k\leq K}$, $(\mu_k)_{1\leq k\leq K}$ and $(C_k)_{1\leq k\leq K}$. We will discuss two cases whether γ is known or not in Section 3.2.3. As the C_k's are symmetric positive definite matrices, $6K$ or $6K+1$ parameters have to be estimated. The aim of this paper is then to estimate these parameters using a set of G independent and partially observed realizations $(\boldsymbol{X}^g, g = 1, \ldots, G)$ of the process (X_t). The $n_g + 1$ observations of the g-th realization are observed at times $(t_0^g, \ldots, t_{n^g}^g)$ and are denoted $(X_0^g, \ldots, X_{n^g}^g)$.
Parameter estimation for diffusion processes is a complicated task due to the unavailability of the transition density of the Markov process defined by (3.23), see Sorensen (2004) and the references therein for a recent survey. Many methodologies introduce an approximation of this transition density, mostly based on Euler scheme, see for instance Aït-Shalia (2008) for an explicit approximation which converges uniformly on the parameter space to the true transition density. See also Durham and Gallant (2002); Pedersen (1995) for methods relying on approximate Monte Carlo maximum likelihood or Elerian *et al.* (2001); Eraker (2001); Roberts and Stramer (2001) for Markov chain Monte Carlo methods combined with data augmentation.
In this paper, the estimation procedure is based on the exact algorithm introduced in Beskos *et al.* (2006a). This algorithm allows to sample skeletons of trajectories exactly distributed as the finite dimensional distribution of the target SDE (3.23). These skeletons can then be used to obtain unbiased Monte Carlo estimates of the intermediate quantity of the Expectation Maximization algorithm to define maximum likelihood estimates of θ, as described in Beskos *et al.* (2006b).

3.2.3 Expectation Maximization based procedure

This section provides an algorithm to estimate the parameter θ using a set of G independent trips $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^G$. Statistical inference based on a set of observations usually requires the finite-dimensional distributions of the process (X_t) to be available. However, in the context of this paper, thanks to the results of Beskos and Roberts G. (2005) and Beskos *et al.* (2006b), the likelihood of the complete path (X_t) with respect to a reference measure on $(\mathsf{C}, \mathcal{C})$ can be obtained easily using Girsanov theorem even if the finite dimensional distribution of $(X_{t_0}, \ldots, X_{t_n})$ is not explicitly available. This setting is conducive to the use of the EM algorithm proposed by Dempster *et al.* (1977) which allows to perform maximum likelihood estimation when the joint distribution of the observations and some additional missing data is available. Starting with an initial estimate θ_0, the EM algorithm produces iteratively a sequence $(\theta_p)_{p\geq 0}$ that converges toward a local maximum of the likelihood of the observations. The EM based algorithm proposed in this paper relies on the Monte Carlo EM procedure presented in Beskos *et al.* (2006b).

E-step

For the sake of clarity, the E-step is first presented assuming that the diffusion γ is known, the general case is developped in the following section.

Case where the diffusion coefficient γ is known : in this case, the parameter to be estimated is

$$\theta := \{(\pi_k)_{1\leq k\leq K}\ ;\ (\mu_k)_{1\leq k\leq K}\ ;\ (C_k)_{1\leq k\leq K}\}\ .$$

For all θ, the law of the process solution to (3.23) is absolutely continuous with respect to the law of the driftless diffusion $\gamma \mathrm{d}W_t$. In this case, the dominating measure on $(\mathsf{C}, \mathcal{C})$ does not depend on θ as γ is known so that maximum likelihood estimation of θ is possible. The procedure proposed in Beskos *et al.* (2006b) is based on rejection sampling and uses a reparametrization of the observations to simplify the dominating measure : the process $(X_t)_{0\leq t\leq T}$ is transformed into a new diffusion process $(Y_t)_{0\leq t\leq T}$ with unit diffusion coefficient. Define the Lamperti transform $\eta : x \mapsto \gamma^{-1}x$. By Ito's formula, the process $(Y_t) = \eta(X_t))_{0\leq t\leq T}$ satisfies $Y_0 = \gamma^{-1}X_0$, $Y_T = \gamma^{-1}X_T$ and

$$\mathrm{d}Y_t = \alpha_\theta(Y_t)\mathrm{d}t + \mathrm{d}W_t\,, \tag{3.28}$$

where α_θ is given by

$$\begin{aligned} &\alpha_\theta : \mathbb{R}^2 \mapsto \mathbb{R}^2 \\ &x \mapsto \gamma^{-1} b_\theta(\gamma x) = -\sum_{k=1}^{K} \pi_k \varphi_k^\theta(\gamma x)\gamma^{-1} C_k^{-1}(\gamma x - \mu_k)\,. \end{aligned} \tag{3.29}$$

Let $\mathbb{Q}_T^\theta$ be the law of $(Y_t)_{0\leq t\leq T}$ on $(\mathsf{C}, \mathcal{C})$ when the SDE is parameterized by θ. By Girsanov formula, $\mathbb{Q}_T^\theta$ is absolutely continuous with respect to the Wiener measure $\mathbb{W}_T$ on $(\mathsf{C}, \mathcal{C})$ and its

Radon-Nikodym derivative is given by :

$$\frac{\mathrm{d}\mathbb{Q}_T^\theta}{\mathrm{d}\mathbb{W}_T}(\omega) = \exp\left\{\int_0^T \alpha_\theta(\omega_s)\mathrm{d}\omega_s - \frac{1}{2}\int_0^T \|\alpha_\theta(\omega_s)\|^2\mathrm{d}s\right\} .$$

Then, applying Ito's formula,

$$\frac{\mathrm{d}\mathbb{Q}_T^\theta}{\mathrm{d}\mathbb{W}_T}(\omega) = \exp\left\{H_\theta(\omega_T) - H_\theta(\omega_0) - \frac{1}{2}\int_0^T \left[\Delta H_\theta(\omega_s) + \|\alpha_\theta(\omega_s)\|^2\right]\mathrm{d}s\right\} ,$$

where H_θ is defined by (3.27) and

$$\Delta H_\theta : x \mapsto \frac{\partial \alpha_{\theta,1}}{\partial x_1}(x) + \frac{\partial \alpha_{\theta,2}}{\partial x_2}(x) . \tag{3.30}$$

Therefore, the log-likelihood function of a complete path is given by :

$$L(\omega,\theta) := H_\theta(\omega_T) - H_\theta(\omega_0) - \frac{1}{2}\int_0^T \left[\Delta H_\theta(\omega_s) + \|\alpha_\theta(\omega_s)\|^2\right]\mathrm{d}s .$$

For each trip $1 \le g \le G$, the observations $\boldsymbol{X}^g$ are transformed into $\boldsymbol{Y}^g$ where for all $0 \le k \le n^g$, $Y_k^g = \gamma^{-1}X_k^g$. Given a current estimate θ_p, the E-step consists in the computation of the intermediate quantity $\theta \mapsto Q(\theta,\theta_p)$ given, as the transformed trips $(\boldsymbol{Y}^g, g = 1,\dots,G)$ are independent, by

$$\begin{aligned}
Q(\theta,\theta_p) &:= \sum_{g=1}^G \mathbb{E}_{\theta_p}\left[L(Y,\theta)|\boldsymbol{Y}^g\right] , \\
&= \sum_{g=1}^G \left\{ H_\theta(Y_{n^g}^g) - H_\theta(Y_0^g) - \frac{1}{2}\mathbb{E}_{\theta_p}\left[\int_0^{t_{n^g}^g} \left[\Delta H_\theta(Y_s) + \|\alpha_\theta(Y_s)\|^2\right]\mathrm{d}s \middle| \boldsymbol{Y}^g\right]\right\} , \\
&= \sum_{g=1}^G \left\{ H_\theta(Y_{n^g}^g) - H_\theta(Y_0^g) - \frac{1}{2}\sum_{j=1}^{n^g}\mathbb{E}_{\theta_p}\left[\int_{t_{j-1}^g}^{t_j^g} \left[\Delta H_\theta(Y_s) + \|\alpha_\theta(Y_s)\|^2\right]\mathrm{d}s \middle| \boldsymbol{Y}^g\right]\right\} ,
\end{aligned}$$

where $\mathbb{E}_{\theta_p}[\cdot|\boldsymbol{Y}^g]$ denotes the conditional expectation under the law of the process (Y_t) on $(\mathsf{C},\mathcal{C})$ given $\boldsymbol{Y}^g$ when the SDE is parameterized by θ_p.

The conditional expectations required to compute the intermediate quantity $\theta \mapsto Q(\theta,\theta_p)$ are not available analytically but, as noted by Beskos *et al.* (2006b),

$$\int_{t_{j-1}^g}^{t_j^g} \left[\Delta H_\theta(Y_s) + \|\alpha_\theta(Y_s)\|^2\right]\mathrm{d}s = \left(t_j^g - t_{j-1}^g\right)\mathbb{E}_{\theta_p}\left[\Delta H_\theta(Y_{U^{g,j}}) + \|\alpha_\theta(Y_{U^{g,j}})\|^2 \middle| (Y_t)_{0\le t\le T}\right] ,$$

where $U^{g,j}$ is independent of $(Y_t)_{0\le t\le T}$ and uniformly distributed on $[t_{j-1}^g, t_j^g]$. Then, the quantity $Q(\theta,\theta_p)$ may be estimated by Monte Carlo simulations. For all $1 \le g \le G$, $1 \le j \le n^g$ and all $1 \le i \le N_j^g$,

(i) simulate $(U_k^{g,j,i})_{1\le k\le M_j}$ independently and uniformly on $[t_{j-1}^g, t_j^g]$;

(ii) conditional on Y_{j-1}^g and Y_j^g, sample a skeleton $Y^{g,j,i}$ at time instances $(U_k^{g,j,i})_{1\le k\le M_j^g}$.

Then, $Q(\theta, \theta_p)$ is estimated by $Q^N(\theta, \theta_p)$ where

$$Q^N(\theta, \theta_p) := \sum_{g=1}^{G} \left[H_\theta(Y^g_{n^g}) - H_\theta(Y^g_0) - \frac{1}{2} \sum_{j=1}^{n^g} \frac{t^g_j - t^g_{j-1}}{M^g_j N^g_j} \sum_{i=1}^{N^g_j} \sum_{k=1}^{M_j} \left\{ \Delta H_\theta \left(Y^{g,j,i}_{U^{g,j,i}_k} \right) + \left\| \alpha_\theta \left(Y^{g,j,i}_{U^{g,j,i}_k} \right) \right\|^2 \right\} \right] .$$

The procedure to sample each $Y^{g,j,i}_{U^{g,j,i}_k}$ given Y^g_{j-1} and Y^g_j is the Exact Algorithm 1 (EA1) of Beskos *et al.* (2006a). It is detailed in Appendix D.1 for completeness along with technical results in Appendix D.2 for the specific implementation details to be applied to the model presented in this paper.

Case where the diffusion coefficient γ is unknown : in this case, the parameter to be estimated is

$$\theta := \{ (\pi_k)_{1 \leq k \leq K} \; ; \; (\mu_k)_{1 \leq k \leq K} \; ; \; (C_k)_{1 \leq k \leq K} \; ; \; \gamma \} \; .$$

Then, the transformation η to obtain the unitary diffusion (3.28) depends on θ. For all $1 \leq g \leq G$, the set $(Y^g_0(\theta), \ldots, Y^g_{n^g}(\theta))$ is not directly observed but is a function of the unknown parameter θ. Beskos *et al.* (2006b) suggested to use a second path transformation to define a dominating measure which does not depend on θ. Define, for all $1 \leq g \leq G$, $1 \leq j \leq n^g$ and all $s \in [t^g_{j-1}, t^g_j]$,

$$\dot{Y}^g_s(\theta) := Y^g_s(\theta) - \left(1 - \frac{s - t^g_{j-1}}{t^g_j - t^g_{j-1}} \right) Y^g_{j-1}(\theta) - \frac{s - t^g_j}{t^g_j - t^g_{j-1}} Y^g_j(\theta) \, . \tag{3.31}$$

The transformation (3.31) maps $(Y^g(\theta))_{t^g_{j-1} \leq s \leq t^g_j}$ onto the diffusion bridge $(\dot{Y}^g)_{t^g_{j-1} \leq s \leq t^g_j}$ starting and ending at 0 for all θ. The law of this transformed process $(\dot{Y}^g)_{t^g_{j-1} \leq s \leq t^g_j}$ is absolutely continuous with respect to the law of the Brownian bridge on $[t^g_{j-1}, t^g_j]$ starting and ending at 0. The inverse transform of (3.31) is given by :

$$f_\theta(\dot{Y}^g_s(\theta')) := \dot{Y}^g_s(\theta') + \left(1 - \frac{s - t^g_{j-1}}{t^g_j - t^g_{j-1}} \right) Y^g_{t^g_{j-1}}(\theta) + \frac{s - t^g_j}{t^g_j - t^g_{j-1}} Y^g_{t^g_j}(\theta) \, . \tag{3.32}$$

The complete path used in the EM algorithm is now $\{ (\dot{Y}^g_t)_{0 \leq t \leq T} \; ; \; \boldsymbol{Y}^g \}$ and the intermediate quantity of the EM algorithm is, by (Beskos *et al.*, 2006b, Lemma 2),

$$\begin{aligned} Q(\theta, \theta_p) = \sum_{g=1}^{G} \Bigg\{ & -2n^g \log \gamma + H_\theta(Y^g_{n^g}(\theta)) - H_\theta(Y^g_0(\theta)) + \sum_{j=1}^{n^g} \log \phi_{t^g_j - t^g_{j-1}}(Y^g_j(\theta) - Y^g_{j-1}(\theta))) \\ & - \frac{1}{2} \sum_{j=1}^{n^g} \mathbb{E}_{\theta_p} \left[\int_{t^g_{j-1}}^{t^g_j} \left[\Delta H_\theta \left(f_\theta(\dot{Y}^g_s(\theta_p)) \right) + \| \alpha_\theta \left(f_\theta(\dot{Y}^g_s(\theta_p)) \right) \|^2 \right] \mathrm{d}s \middle| \boldsymbol{Y}^g \right] \Bigg\} \, , \end{aligned}$$

where ϕ_u is the probability density function of a 2 dimensional $\mathcal{N}(0, uI)$ random variable. Then, following the same steps as in the case where γ is known, the quantity $Q(\theta, \theta_p)$ may be estimated by Monte Carlo simulations. For all $1 \leq g \leq G$, $1 \leq j \leq n^g$ and all $1 \leq i \leq N^g_j$,

(i) simulate $(U_k^{g,j,i})_{1\leq k\leq M_j^g}$ independently and uniformly on $[t_{j-1}^g, t_j^g]$;

(ii) conditional on $Y_{j-1}^g(\theta_p)$ and $Y_j^g(\theta_p)$, draw a skeleton $Y^{g,j,i}$ at times $(U_k^{g,j,i})_{1\leq k\leq M_j^g}$;

(iii) compute $\dot{Y}^{g,j,i}(\theta_p)$ at time instances $(U_k^{g,j,i})_{1\leq k\leq M_j^g}$ by evaluating (3.31) at $Y^{g,j,i}_{U_k^{g,j,i}}(\theta_p)$.

Then, $Q(\theta, \theta_p)$ is estimated by $Q^N(\theta, \theta_p)$ where

$$Q^N(\theta,\theta_p) := \sum_{g=1}^{G}\left\{-2n^g\log\gamma + \sum_{j=1}^{n^g}\log\phi_{t_j^g-t_{j-1}^g}\left(Y_j^g(\theta)-Y_{j-1}^g(\theta)\right) + H_\theta(Y_{n^g}^g) - H_\theta(Y_0^g)\right.$$
$$\left. - \frac{1}{2}\sum_{j=1}^{n^g}\frac{t_j^g - t_{j-1}^g}{M_j^g N_j^g}\sum_{i=1}^{N_j^g}\sum_{k=1}^{M_j^g}\left\{\Delta H_\theta\left(f_\theta\left(\dot{Y}^{g,j,i}_{U_k^{g,j,i}}\right)\right) + \left\|\alpha_\theta\left(f_\theta\left(\dot{Y}^{g,j,i}_{U_k^{g,j,i}}\right)\right)\right\|^2\right\}\right\} . \tag{3.33}$$

M-step

In both cases (γ known or unknown), as the function $\theta \mapsto Q^N(\theta, \theta_p)$ cannot be maximized analytically, the M-step is performed numerically. This step is based on the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) introduced in Hansen and Ostermeier (2001) which is a derivative-free optimization procedure. The CMA-ES is known to perform well for complex multimodal optimization problems, see e.g. Hansen and Kern (2004). In our case, the constrained version of CMAES should be used but the following parametrization circumvents the constraints problem and the classical version of CMA-ES algorithm is finally used :

- each C_k is a positive definite matrix and may be written

$$C_k = \begin{pmatrix} \exp(2a_1^k) & a_3^k\exp(a_1^k) \\ a_3^k\exp(a_1^k) & \exp(2a_2^k) + (a_3^k)^2 \end{pmatrix} ,$$

  where $a_1^k, a_2^k, a_3^k \in \mathbb{R}$;

- for all $1 \leq k \leq K$, $\pi_k > 0$ and may be written $\exp(\tilde{\pi}_k)$, with $\tilde{\pi}_k \in \mathbb{R}$;

- $\gamma > 0$ and may be written $\exp(\tilde{\gamma})$, with $\tilde{\gamma} \in \mathbb{R}$.

Parameters of the CMA-ES algorithm are tuned according to the heuristics given in Hansen and Kern (2004), except for the initial standard deviation at each MCEM step. It is chosen to be increasing from a small to a large value during the first iterations, and then decreasing from this large value to a smaller one for last iterations. This method allows a good exploration of the parameter space without using time consuming adaptative techniques.

Approximate log likelihood

The EM algorithm is known to ensure that, for a given starting point θ_0 the sequence $(\theta_p)_{p\in\mathbb{N}}$ converges towards a local maximum of the likelihood, say $\hat{\theta}_1$. However, if for another starting point, another local maximum was reached, say $\hat{\theta}_2$, there is a need to compare the likelihood of both parameters to choose the best one.

The estimation approach adopted in this paper provides, for a given estimator $\hat{\theta}$, a surrogate for the loglikelihood of the observed paths by computing $Q(\hat{\theta}, \hat{\theta})$ using equation (3.33).
The log likelihood is given by

$$\sum_{g=1}^{G} L(\mathbf{Y}^g, \hat{\theta}),$$

where $L(\mathbf{Y}^g, \hat{\theta})$ is the loglikelihood of the g-th observed path. However, it can't be computed as

$$\sum_{g=1}^{G} L(\mathbf{Y}^g, \hat{\theta}) = Q(\hat{\theta}, \hat{\theta}) - \sum_{g=1}^{G} \mathbb{E}_{\hat{\theta}} \left[\log p_{\hat{\theta}}(\mathbf{Y}^{-g} | \mathbf{Y}^g) \middle| \mathbf{Y}^g \right] ,$$

where $p_{\hat{\theta}}(\mathbf{Y}^{-g}|\mathbf{Y}^g)$ is the conditional distribution of the missing path given $\mathbf{Y}^g$ when the parameter is $\hat{\theta}$ which is not available analytically.
Nevertheless, following (Beskos *et al.*, 2009, Theorem 1), the properties of α_θ (equation (3.29)) allows us to use Monte Carlo simulations to obtain an estimator $L^N(\mathbf{Y}^g, \hat{\theta})$ of the loglikelihood $L(\mathbf{Y}^g, \hat{\theta})$ for all $1 \leq g \leq G$ (N refers the to the numbers of of simulated particles). This estimator is derived on appendix D.3. Therefore, for any θ_1 and θ_2 resulting from the MCEM procedure, we are able to choose the one that gives the best likelihood.

Model Selection

In the previous section, the number of mixture components K has been assumed to be known and fixed. In the context of mixture models, following Biernacki *et al.* (2000) the number of components is classically selected according to Integrated Completed Likelihood (ICL) criterion. Even if the shape of the target potential function reminds this context, the missing data in the presented work are the full trajectories and not the component identifier, therefore the ICL citerion can not be easily derived in our context. The number of components is obtained by approximating the Akaïke Information criterion by :

$$AIC(\hat{\theta}) = -2 \sum_{g=1}^{G} L(\mathbf{Y}^g, \hat{\theta}) + 2\dim(\hat{\theta}) \,, \tag{3.34}$$

As we seen on the above section, the likelihood is unknown here. However, we can derive an approximation of the AIC criterion by using the estimator of the log likelihood $L^N(\mathbf{Y}^g, \hat{\theta})$. The approximate AIC criterion is therefore

$$\hat{AIC}(\hat{\theta}) = -2 \sum_{g=1}^{G} L^N(\mathbf{Y}^g, \hat{\theta}) + 2\dim(\hat{\theta}) \,, \tag{3.35}$$

3.2.4 Experimental results

Simulated data set

This section illustrates the performance of our procedure using simulated data in the case where the diffusion coefficient γ is unknown. For a given set of parameters, a toy set of trips

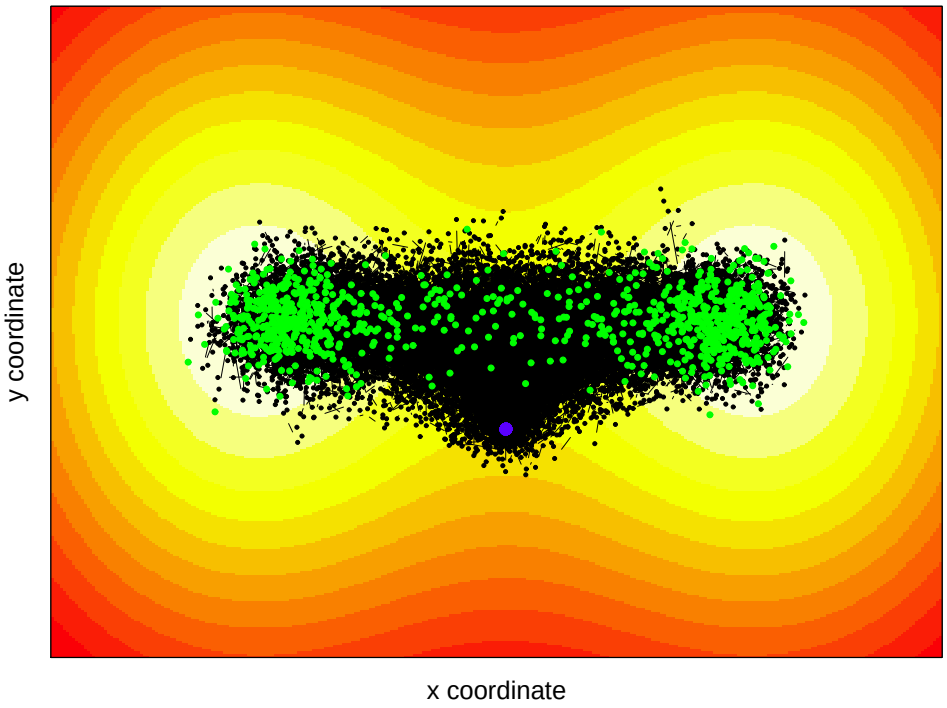


FIGURE *3.6 – 1000 simulated trajectories for $K = 2$ starting from a unique X_0 (blue point). Each trajectory is sampled 30 times. Final points are colored in green. High (resp. low) potential areas of the underlying map are represented in white (resp. red).*

$\boldsymbol{X}^1, \ldots, \boldsymbol{X}^G$ satisfying (3.23) is simulated using the exact algorithm EA1 of Beskos *et al.* (2006a). From this data set, the MLE is estimated using the algorithm described in section 3.2.3. The estimation is performed with $K = 1$ (7 parameters to estimate) and $K = 2$ (14 parameters to estimate). In each case, 3 configurations are tested :

- **Scenario 1 :** G trajectories starting from G different X_0^g, with n observations sampled at regular times ;
- **Scenario 2 :** G trajectories starting from a unique X_0, with n observations sampled at regular times (time step set at $\delta = 0.25$) ;
- **Scenario 3 :** G trajectories starting from a unique X_0, with n observations sampled at irregular times. Trajectories with $5n$ points are simulated with a regular time step $\delta = 0.05$, and n observations are drawn uniformly from these trajectories.

The first configuration is an "ideal case" as the exploration of the space is better with different starting points. The second and third configurations are more likely to happen in a real context since tagged individuals often start there trips from a unique point such as a colony or a harbor. Moreover, in ecology, most of timesteps acquisition of GPS positions are irregular due to environmental conditions. Examples of simulated trajectories for ($K = 2$, scenario 2) are shown in Figure 3.6. For $K = 1$ and $K = 2$, we set $G = 500$, $n = 31$, corresponding to numerous but short trajectories. For each sampling scheme, the MCEM algorithm is performed from 50 starting points $\theta_0^{(i)}, i = 1, \ldots, 50$ giving 50 estimates $\hat{\theta}^{(i)}, i = 1, \ldots, 50$. Then, the final estimate

is given by

$$\hat{\theta} = \text{argmax}_i \sum_{g=1}^{G} L^N(\mathbf{Y}^g, \hat{\theta}^{(i)}),$$

where $L^N(\mathbf{Y}^g, \hat{\theta}^{(i)})$ is the estimator of the loglikelihood $L(\mathbf{Y}^g, \hat{\theta}^{(i)})$ of the g-th trajectory (see Appendix D.3). The number of Monte Carlo samples to approximate the expectation increases with EM iterations (following Fort and Moulines (2003)), up to a maximum of 100 particles per segment.

Choosing the starting point θ_0 of EM algorithm A key aspect in the behaviour of the EM algorithm deals with the initial set of parameters : choosing an appropriate θ_0 is crucial to design a time efficient estimation procedure. We consider some heuristics to pick a first guess, that would, of course, depend on the experiment.

- First guess for μ and C may be chosen using Gaussian mixture estimators, ignoring the correlation between successive relocations. However, this technique requires a subsampling of trajectories to get rid of the autocorrelation of the observed processes. This method may also be used to get the relative weight of each attractive zone.
- Choosing a good first guess for γ may be done using different estimators for diffusion process, for instance using Brownian bridges techniques. In practice, the estimation of this parameter is hardly sensitive to the starting point.
- A first guess for $\sum \pi_k$ may be trickier to find as it is strongly related to the speed of the individuals and thus requires expert knowledge of the experimental setting. It might be set to one.

Several values are drawn around this heuristically chosen starting set of parameters, corresponding to several trajectories of convergence.

Results An example of MCEM trajectories for ($K = 1$, scenario 3) is shown in Figure 3.7. Map estimates for $K = 1$ and $K = 2$ are presented in Figures 3.8 and 3.9. Detailed values for the best estimates are shown in Tables 3.1 and 3.2. As general comments, for all scenarios, there is a very fast convergence for the parameter γ. In these simple scenarios, our algorithm provides efficient estimates for the location parameters μ. The convergence is often slower for weights parameters π_k and shape parameters C_k. As expected, best estimates are obtained when the sampling is regular and the space is well explored (Scenario 1). Good estimation behavior is still observed when only one starting point is considered, and irregularity in sampling seems to have no impact on the performance of the estimation. This last point might be of great importance when dealing with actual data, as environmental conditions often lead to irregular sampling.

Real data set

GPS positions of a French fishing vessel performing in the English Channel were recorded during one year. The data set consists of 57 trajectories (assumed to be independent). Each

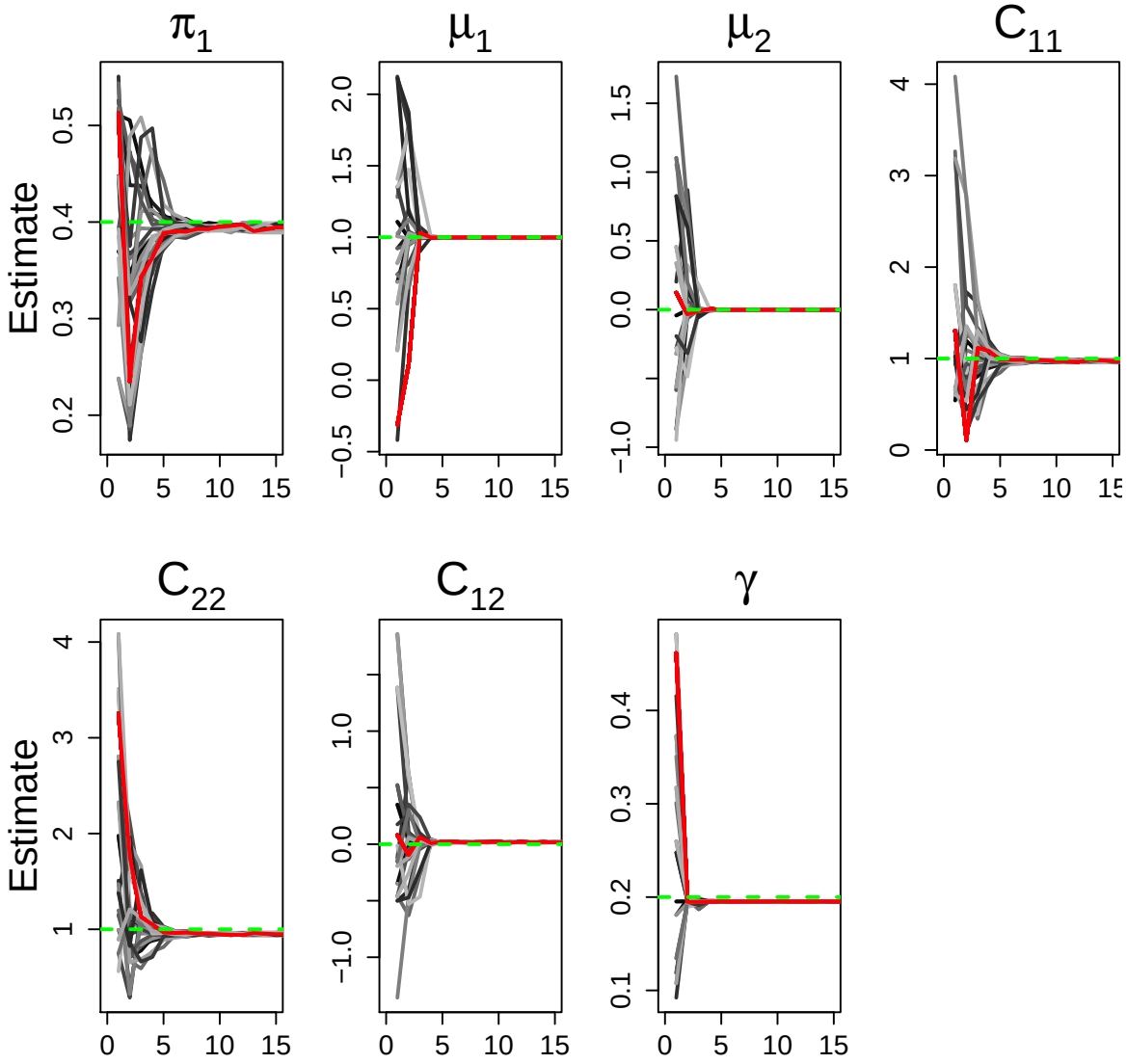


FIGURE *3.7 – Example of MCEM trajectories for ($K = 1$, scenario 3). The red line is the best estimate and the green dotted line is the true value.*

Scenario/Parameter	π_1	$\mu^{(1)}$	$C^{(1)}$	γ
True Value	0.4	$(1, 0)'$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	0.2
Scen. 1	0.394	$(0.999, 0)'$	$\begin{pmatrix} 0.974 & 0.017 \\ 0.017 & 0.951 \end{pmatrix}$	0.196
Scen. 2	0.39	$(1.019, -0.002)'$	$\begin{pmatrix} 0.931 & -0.015 \\ -0.015 & 1.006 \end{pmatrix}$	0.196
Scen. 3	0.388	$(1.046, -0.008)'$	$\begin{pmatrix} 1.013 & -0.113 \\ -0.113 & 1.054 \end{pmatrix}$	0.197

TABLE *3.1 – Best estimates for each scenario when $K = 1$. Results are rounded to 10^{-3}.*

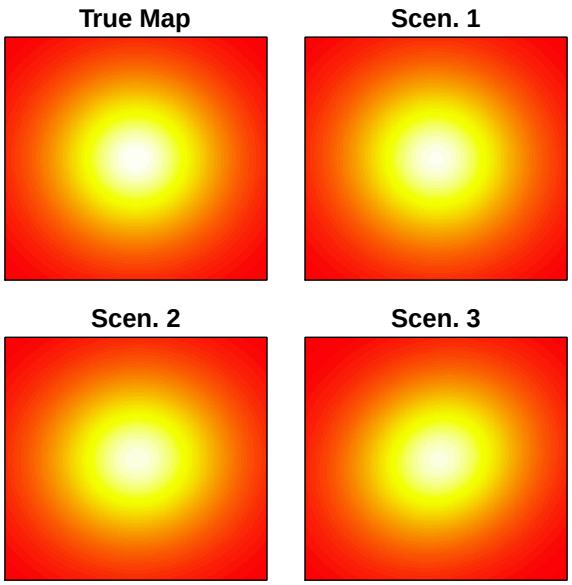


FIGURE *3.8 – Estimated map for the three scenarios when $K = 1$. The scale is the same on each graph.*

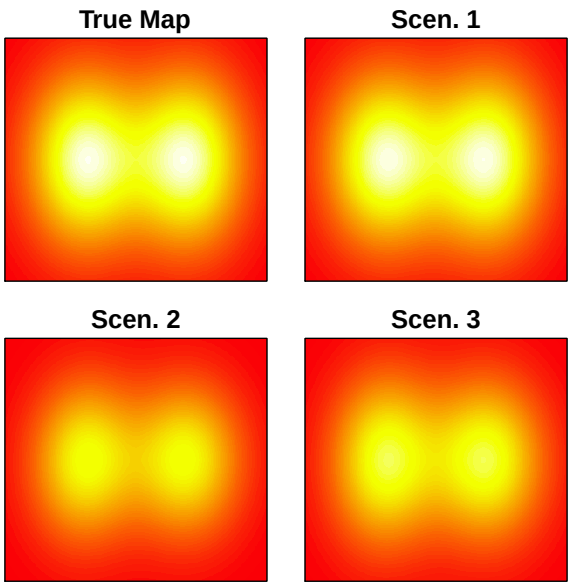
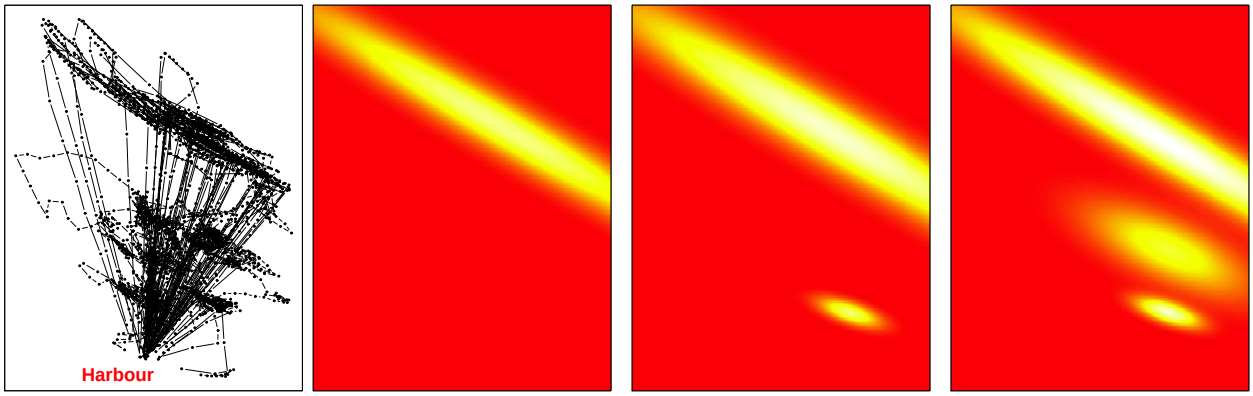


FIGURE *3.9 – Estimated map for the three scenarios when $K = 2$. The scale is the same on each graph.*

Scenario/Parameter	$\pi_1\,,\ \pi_2$	$\mu^{(1)}\,,\ \mu^{(2)}$
True Value	$0.5\,,\, 0.5$	$(1,0)'\,,\ (-1,0)'$
Scen 1.	$0.498\,,\ \ 0.497$	$(1.01,0)'\,,\ (-1.01,0.002)'$
Scen 2.	$0.399\,,\ \ 0.399$	$(0.969,-0.002)'\,,\ (-0.981,-0.005)'$
Scen 3.	$0.434\,,\ \ 0.430$	$(0.974,-0.005)'\,,\ (-0.991,-0.003)'$

Scenario/Parameter	$C^{(1)}\,, C^{(2)}$	γ
True Value	$\begin{pmatrix}1.667 & 0\\ 0 & 1.667\end{pmatrix}\,,\ \begin{pmatrix}1.667 & 0\\ 0 & 1.667\end{pmatrix}$	0.1
Scen 1.	$\begin{pmatrix}1.644 & -0.012\\ -0.012 & 1.633\end{pmatrix}\,,\ \begin{pmatrix}1.628 & -0.007\\ -0.007 & 1.684\end{pmatrix}$	0.098
Scen 2.	$\begin{pmatrix}1.825 & -0.017\\ -0.017 & 2.241\end{pmatrix}\,,\ \begin{pmatrix}1.943 & -0.055\\ -0.055 & 2.059\end{pmatrix}$	0.098
Scen 3.	$\begin{pmatrix}1.691 & -0.002\\ -0.002 & 2.004\end{pmatrix}\,,\ \begin{pmatrix}1.920 & -0.039\\ -0.039 & 1.829\end{pmatrix}$	0.097

TABLE *3.2 – Best estimates for each scenario when* $K = 2$*. Results are rounded to* 10^{-3}*.*



(a) Raw data *(b) $K = 1$* *(c) $K = 2$* *(d) $K = 3$*

FIGURE *3.10 – Estimated attractive zones for a French fishing vessel performing in the Channel. The model presented above is fitted with 1, 2 and 3 modes.*

trajectory is sampled regularly every 15 minutes in average, with about 40 points per trajectory, giving a total of 2723 points. The raw data set is shown on Figure 3.10a. Model (3.23) described above is fitted to this data set with $K = 1$, $K = 2$ and $K = 3$ modes [23]. Starting points for the EM algorithm were chosen randomly, using uniform distributions for position parameters (μ_k) weight parameters (π_k) and diffusion parameter γ, and Wishart distributions for shape parameters (C_k). A hundred iterations of the MCEM algorithm were performed for each point. For this data set, the model with three modes was selected using the approximated AIC criterion. The estimated maps for all values of K are given on figure 3.10. Three non connected zones are identified, a highly attractive deep sea zone, an attractive coastal zone, and an intermediate zone between two waters. The zones are oriented East-West, reflecting the main surface currents directions in the Channel.

3.2.5 Conclusions

This work proposes a new parametric model continuous in time and continuous in space to analyze trajectory data in ecology. This model assumes the trajectory process of an individual to be the realization of the solution to a SDE. The drift of this SDE reflects the attractiveness of the environment. More formally, the drift is defined as the gradient of a potential map. The model is estimated using a Monte Carlo EM algorithm. The Monte Carlo aspects rely on the exact algorithm for simulation proposed by Beskos *et al.* (2006a). The exact simulation algorithm avoids the estimation bias due to discretization schemes, which may be large when the sampling frequency is not high enough. Here, the estimation of the potential map does not require any assumption on the sampling frequency, as would require kernel methods or discrete time models. The potential map is chosen here to be a mixture of positive Gaussian shaped functions. The parametric form of $P(\cdot)$ allows a flexible form for maps, however, the number of parameters increases linearly with the number of modes. A minor change in this work would be to modify the parametric form of $P(\cdot)$ to describe other shapes for attractive areas, keeping

23. The model was coded and fitted using the **R** software R Core Team (2014) coupled with C++, using the package **Rcpp** Eddelbuettel and François (2011). All codes are available on demand.

only the boundary and C^2 properties. Such a change would only require a new computation of technical bounds required for the simulation procedure.

As in any mixture model, the model selection problem arises and should be investigated in our framework. Thanks to Beskos *et al.* (2009), an approximation of the loglikelihood can here be computed for any parameter θ, therefore a penalized likelihood based criterion can be used. We used here an approximation of the AIC criterion. The approximation of the AIC proposed here is different from the one proposed by Uchida and Yoshida (2005), that assumed the process solution to the SDE to be ergodic, as this criterion is mostly computed using one unique realization, observed for a long time. This criterions couldn't theoretically be used in our case as the solution of the proposed SDE is not ergodic.

The process considered in equation (3.23) exhibits Brownian Motion like behaviour when far from the attractive zone, and is therefore non ergodic. Nevertheless good asymptotic properties are obtained as the number of trajectories G considered goes to infinity and the property of ergodicity is no more required. Since this model is intended to be used for ecological trajectory data sets and the structure of the data will often consist in a large amount of (presumably) independent trajectories that allows the state space to be visited. The ergodic property for the solution of our process would be satisfied at the cost of a change in the drift adding a term which avoids the process to visit space to far from the attractive points.

This model assumes that the diffusion coefficient is scalar, which is mainly a technical constraint. Using the approach presented here, the diffusion must satisfy a conservative property, that is actually the core of our model. A major limitation of our approach for ecological users might be the time homogeneous potential map. This assumption guarantees the process solution to the SDE (3.23) to have Markovian properties which greatly simplifies the estimation procedure. However, it is known that an individual might adopt different behaviors during its trips. It might be interesting to introduce state space models where different behaviors are considered.

Another interesting improvement would be to add interactions between individuals. An attraction/repulsion term could be added in the drift to indicate whether individuals are likely to cooperate or not. Another limitation of our process is the non existence of a stationary distribution which is linked to the non ergodic property mentionned before. This is not a problem in practice since the starting point is close to attractive zones but, in theory, ecologists would rather consider an additional term to ensure the existence of a stationary distribution. Thsi would be mandatory if considering only one unique long trajectory. As said above, if this term still allow the function in (3.29) to satisfy the conservative property, the estimation framework presented here would remain valid.

To conclude, we believe the model presented here offers new insight for trajectory analysis.

We propose a general model continuous in time and space, that requires no assumption on sampling frequency. We think that this gradient based model and its estimation framework could be easily extended to answer many ecological questions.

3.3 Application du modèle GaP

Cette section propose un exemple d'application du modèle GaP présenté ci-dessus. On se propose de répondre à la question suivante : La perception estimée de la distribution de la seiche en Manche Est au travers de l'activité des pêcheurs est elle corrélée à sa distribution estimée par la campagne scientifique ?

3.3.1 Introduction

La structure de répartition des proies est un facteur déterminant du comportement des prédateurs (Benoit-Bird and Au, 2003, par exemple). Pour se nourrir, un prédateur a tout intérêt à concentrer sa prospection là où la ressource est abondante. Ainsi, les déplacements d'un prédateur peuvent renseigner sur la distribution spatiale de la ressource ciblée. Plusieurs modèles ont été développés pour décrire la manière dont se répartissent des prédateurs en fonction de leurs proies. Le plus célèbre, et qui est à la base de nombreux autres modèles est le modèle de distribution idéale libre ("Ideal Free Distribution", IFD Fretwell (1972)), stipulant que les prédateurs sont égaux entre eux, omniscients et rationnels. Appliqué à la pêche, ce modèle suppose que l'activité des navires de pêche serait un parfait révélateur de la ressource ciblée. Cette hypothèse a été testée sur la pêcherie canadienne de crabe des neiges, montrant que l'activité de pêche était un bon proxy de l'abondance (Swain and Wade, 2003). Le modèle IFD est également souvent utilisé pour modéliser le comportement des pêcheurs, car il nécessite moins d'hypothèses que les modèles de décision probabilistes (Gillis, 2003). Ces études ont cependant été faites à une échelle temporelle large, et souvent avec une agrégation spatiale grossière des observations.

Grâce à leur résolution spatiale fine, les données VMS, offrent la possibilité de tester à une échelle fine l'hypothèse de la distribution idéale libre. En effet, l'accès aux trajectoires des pêcheurs tout au long de l'année à un pas de temps fin, peut permettre un croisement temporel et spatial avec les observations scientifiques ayant lieu sur une zone, et à une période particulière.

> *« Studying VMS and [scientific survey] data could be like looking at fish through different glasses : fishermen and scientific perspectives. »*
>
> **Joo (2013)**

La concordance spatiale et temporelle des deux activités, de pêche, et d'observation scientifique, donne deux points de vue différents sur une même problématique. L'activité de pêche est souvent qualifiée d'échantillonnage préférentiel de la ressource. En effet, l'activité de pêche a pour but d'être rentable, elle devrait donc, en toute logique, se diriger vers les concentrations de ressource les plus rentables. L'activité scientifique, répond à un échantillonnage systématique d'une zone, et en ce sens, doit échantillonner une zone de manière plus complète et objective. Il est par conséquent naturel de s'attendre à une corrélation positive entre les deux perceptions de la ressource ciblée par les pêcheurs. Une telle analyse a été proposée par Joo (2013) sur la pêcherie d'anchois du Pérou. Cependant, cette analyse n'a pu montrer de corrélation claire

entre les deux perceptions.
Dans cette l'étude de Joo (2013), les auteurs ont estimé une activité pour chaque position observée le long des trajectoires de navires, à l'aide d'un modèle de semi-Markov caché [24]. Cette activité pouvait être la pêche, la route ou la recherche. Une carte de probabilité de présence de la ressource était ensuite estimée en faisant les hypothèses suivantes :

- Pour une position en pêche, la probabilité de présence de la ressource au même instant à cette position est égale à 1 ;
- Pour une position en route, la probabilité de présence de la ressource au même instant à cette position est égale à 0 ;
- Pour une position en recherche, la probabilité de présence de la ressource au même instant à cette position est calculée comme étant une fonction dépendant de plusieurs facteurs.

Une carte de probabilité de présence de la ressource peut alors être obtenue et comparée à la carte d'abondance estimée à partir des campagnes acoustiques effectuées dans la même zone. Aucune corrélation significative n'a pu être mise en évidence.

Nous proposons ici une approche différente. En se basant sur le modèle de mouvement développé dans la section précédente, nous partons du principe que les trajectoires des navires de pêche révèlent une perception subjective de la ressource ciblée. Le but est de comparer ce champ subjectif, estimé à partir des trajectoires, à un champ (censé être) objectif, estimé à partir d'un campagne scientifique. Cette approche part du principe que le mouvement des navires de pêche est la conséquence d'une connaissance subjective de l'attractivité de l'espace, et que cette perception est fixe sur une période donnée [25] (celle où a lieu la campagne scientifique). L'hypothèse sous-jacente est que cette attractivité est un reflet de l'abondance de la ressource.

Nous nous intéressons à la pêcherie de la seiche en Manche Est pendant le mois d'Octobre, mois durant lequel se déroule la campagne scientifique Channel Ground Fish Survey (CGFS), ayant pour but, entre autres, l'aide à l'évaluation du stock de la seiche. Nous cherchons alors à répondre à la question : La perception estimée de la distribution de la seiche en Manche Est au travers de l'activité des pêcheurs est elle corrélée à sa distribution estimée par la campagne scientifique ?

3.3.2 Données

Les données VMS

On dispose des enregistrements VMS des navires de pêche français exerçant dans la Manche Est. Ces données consistent en des points GPS acquis en moyenne toutes les heures, de manière plus ou moins régulière (tableau 3.3).

24. Reprenant les idées du chapitre 2
25. Cette condition traduit le fait que le modèle GaP présenté a une fonction de dérive homogène en temps

	Baie de Seine (2012)	Boulogne (2008)
Nombre de trajectoires	48	41
Nombre de points	1784	2443
$\bar{\Delta}t$ (écart type)	62mn (21mn)	53mn (34mn)

TABLE *3.3 – Caractéristiques des données VMS pour les deux jeux de données. $\bar{\Delta}t$ désigne la moyenne du pas de temps d'acquisition sur les trajectoires considérées.*

Dans cette étude, on ne considère que les marées [26] ayant eu lieu en Octobre, ciblant la seiche, et ayant lieu dans deux zones distinctes. Plus précisément, trois filtres sont ainsi appliqués sur les trajectoires :

- Un filtre temporel. On ne conserve que les trajectoires ayant eu lieu dans le courant du mois d'Octobre. On s'assure ainsi de la concordance entre la période d'activité du navire et la période de l'échantillonnage scientifique.
- Un filtre sur les espèces capturées. En se basant sur les données déclaratives de débarquement, on ne considère que les trajectoires où l'espèce capturée en majorité était la seiche. En appliquant ce filtre, on se concentre ainsi sur les trajectoires où l'espèce ciblée est la seiche. On a ainsi une concordance entre l'espèce cible des pêcheurs et l'espèce observée par les scientifiques.
- Un filtre zone. Deux zones sont considérées, correspondant aux zones d'exploitation de deux des ports les plus importants de la Manche Est :
  - Zone Z_1 : La Baie de Seine, zone d'exploitation dont le port H_1 dans cette étude est Port-en-Bessin.
  - Zone Z_2 : Le large de Boulogne sur Mer, zone d'exploitation dont le port H_2 dans cette étude est Boulogne sur Mer.

  Ces deux zones sont exploitées généralement de manières distinctes par les navires ciblant la seiche (figure 3.11).

Dans cette étude, on se focalise sur une année par zone, chaque année étant choisie pour disposer d'un nombre conséquent de marées ciblant la seiche, permettant ainsi l'estimation de carte de perception à partir de ces données (l'année 2008 pour Boulogne, et 2012 pour la Baie de Seine). Les caractéristiques des données VMS pour ces deux jeux de trajectoires sont résumées dans le tableau 3.3

La campagne CGFS

La campagne scientifique Channel Ground Fish Survey (CGFS) a pour objectif de "collecter les données de base pour une estimation de l'état des ressources, par une évaluation directe de l'abondance des stocks et de leur distribution, associée à l'échantillonnage biologique des captures" [27]. Elle est effectuée chaque année depuis 1988 pendant le mois d'Octobre. Elle couvre toute la Manche Est (zone CIEM VIId) avec un plan d'échantillonnage systématique (figure

26. Une marée est la sortie en mer comprise entre un départ d'un port, et au retour à un port, potentiellement distinct

27. `wwz.ifremer.fr/manchemerdunord/Unite-Halieutique/Halieutique-Boulogne-sur-Mer/Axes-de-recherche/Dynamique-des-pecheries/Campagnes-scientifiques/CGFS`

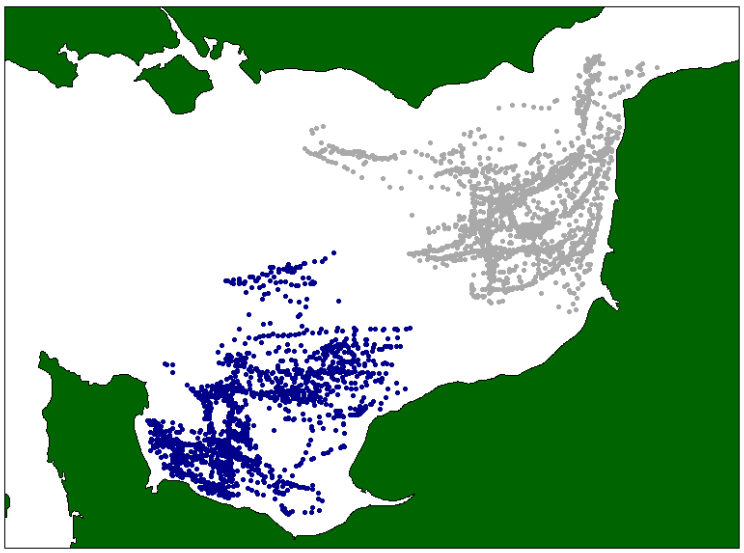


FIGURE *3.11 – Exemples de trajectoires de navires de pêche ciblant la seiche en Octobre, pour deux zones : La Baie de Seine (bleu), et Boulogne sur Mer (gris). Les zones d'activités sont distinctes.*

3.12) consistant en des traits de chaluts effectués aux points du plan d'échantillonnage. Dans cette étude, nous nous intéressons en particulier à la seiche.

Pour chaque point d'échantillonnage G de la campagne CGFS, le nombre d'individu et les tailles sont renseignées. On obtient ainsi la masse totale $M(G)$ en seiche du trait de chalut, grâce à une relation taille-poids. Les paramètres de cette relation sont obtenus dans la littérature (Dunn, 1999). De plus, pour chaque trait de chalut, l'aire totale chalutée $A(G)$ est connue, et permet de ramener la capture à une densité.

Afin de circonscrire l'analyse de corrélation à des zones observées par la pêche et par la campagne scientifique, certains points de l'échantillonnage CGFS ne sont pas comparés avec l'activité de pêche correspondante en ces points. À savoir :

- Les points de l'échantillonnage situés dans des zones où la pêche est interdite pour les navires étudiés ici, qui sont :
  - Les points le long de la côte anglaise ;
  - Les points le long de la côte française dans la limite des 3 miles nautiques, où l'activité de chalutage est interdite.
- Les points situés en dehors du champ d'action des navires d'une zone. Le champ d'action est ici défini comme le plus petit rectangle contenant toutes les positions de pêche.

Ces points sont cependant conservés pour une analyse préliminaire, qui est l'estimation de la matrice Σ de l'équation 3.39, car ils donnent une information sur la structure de corrélation spatiale de l'abondance en seiche (paragraphe **Comparaison des deux indices**).

FIGURE *3.12 – Points de l'échantillonnage systématique CGFS dans la Manche Est*

3.3.3 Méthodes

L'indice d'abondance I_{CGFS}

Pour chaque point d'échantillonnage G, de coordonnées (x_G, y_G) on obtient donc un indice d'estimation de l'abondance de seiche sur la zone chalutée

$$I_{CGFS}(G) = \log(M(G)/A(G)) \tag{3.36}$$

Le passage au logarithme permet de lisser les valeurs des indices, au vu de la grande variabilité dans la quantité de seiche obtenue. On notera que, du fait de cette transformation logarithmique, une variation de l'indice I_{CGFS} construit en (3.36) s'interprète comme une variation de la densité de seiche en terme d'ordre de grandeur.
Pour chaque zone Z_i, on dispose d'un nombre des points d'échantillonnage n_{Z_i}. On obtient donc en sortie, deux vecteurs

$$\mathbf{I}^{(\mathbf{i})}_{\mathbf{CGFS}} = (I_{CGFS}(G_1^{(i)}), \ldots, I_{CGFS}(G_{n_{Z_i}}^{(i)})),$$

où $i = 1, 2$ représente la zone considérée.

L'indice de préférence I_{VMS}

Dans cette étude, on considère que les trajectoires de navires de pêche d'une même zone satisfont le modèle de déplacement GaP. Pour une zone Z_i une trajectoire $(X_t)_{0\leq t}$ partant du port H_i est considérée comme un processus continu, solution de l'équation différentielle stochastique

$$X_0 = H_i, \quad dX_t = \nabla P_i(X_t, \theta_i)dt + \gamma_i dW_t \tag{3.37}$$

où $i = 1, 2$ est le numéro de la zone, ∇ est l'opérateur gradient, P est une fonction de potentiel $\mathbb{R}^2 \mapsto \mathbb{R}$, θ_i est l'ensemble des paramètres de cette fonction de potentiel pour la zone Z_i, et γ_i est un paramètre de diffusion (voir la section 3.2.2).
Ainsi, les N_{Z_i} trajectoires des navires dans la zone Z_i sont supposées être N_{Z_i} réalisations indépendantes du processus solution de (3.37).
La fonction P représente le potentiel guidant le mouvement dans la zone Z_i. On interprète $P(x, \theta_i)$ comme la manière dont les pêcheurs perçoivent l'attractivité du point x de l'espace. Pour un point x quelconque de $\mathbb{R}^2$, on définit alors directement l'indice $I_{VMS}(x)$ comme la valeur de P en x :

$$I_{VMS}^{(i)}(x) = P(x, \theta_i).$$

Estimation de l'indice I_{VMS}

La fonction $P(\cdot, \theta)$ est supposée être un mélange de K formes Gaussiennes, telle que définie dans la section 3.2.2 (équation (3.24)). On pose ainsi

$$P(x, \theta) := \sum_{k=1}^{K} \pi_k \varphi_k^{\theta}(x)\,, \tag{3.38}$$
$$\varphi_k^{\theta}(x) := \exp\left(-\frac{1}{2}(x - \mu_k)^T C_k (x - \mu)\right).$$

où $\pi_k \in \mathbb{R}, \mu_k \in \mathbb{R}^2$, C_k une matrice de covariance et K est un nombre entier positif. Le paramètre θ comprend ainsi les paramètres $(\mu_k)_{1 \le k \le K}$, $(C_k)_{1 \le k \le K}$ and $(\pi_k)_{1 \le k \le K}$, soient $6K$ paramètres.

- K représente le nombre de sous-zones attractives composant une zone de pêche ;
- μ_k représente la position du centre de la k-ième sous-zone attractive ;
- C_k représente la forme de la k-ième sous-zone attractive ;
- π_k représente le poids de la k-ième sous-zone attractive dans le mélange.

La section 3.2.3 décrit, pour K connu, la méthode d'estimation du maximum de vraisemblance du couple (θ, γ) en utilisant une procédure MCEM (Fort and Moulines, 2003). L'algorithme est une procédure itérative nécessitant le choix d'un point de départ (θ_0, γ_0) pour le paramètre à estimer. À chaque itération, ce paramètre est actualisé selon une procédure détaillée dans la section 3.2.3. Cette procédure nécessite une approximation Monte Carlo. Cette étape est effectuée en utilisant de 15 à 35 particules par segment de trajectoires (le nombre augmente avec les itérations).

Les procédures de type EM sont dépendantes du point initial choisi. Un point initial est ici un ensemble de valeurs données à tous les paramètres du modèle. Les points initiaux sont choisis aléatoirement en tirant des valeurs de paramètres des lois uniformes (respectivement, Wishart, uniformes, uniformes) pour les paramètres μ_k (respectivement C_k, π_k, γ). Le choix des meilleurs paramètres estimés à partir de ces points de départ est fait en utilisant une ap-

proximation de la vraisemblance du paramètre (calculée grâce à l'équation (D.20)).

Une difficulté réside ici dans le choix du nombre de sous-zones attractives K. Dans cette étude, nous avons séquentiellement ajusté sept modèles comprenant respectivement 1 à 7 sous-zones (K=1,..., 7). Le modèle sélectionné est alors celui minimisant l'approximation du critère AIC (Akaike, 1992) définie dans l'équation (3.35).

La carte estimée $\hat{P}$ étant une fonction continue de l'espace, on calcule pour chaque zone Z_i, et chaque point de l'échantillonnage CGFS, la valeur de l'indice I_{VMS} en ce point. On obtient ainsi, pour chaque zone Z_i, un vecteur

$$\mathbf{I}^{(\mathbf{i})}_{\mathbf{VMS}} = (I_{VMS}(G_1^{(i)}), \ldots, I_{VMS}(G_{n_{Z_i}}^{(i)})).$$

Comparaison des deux indices

Pour tester s'il existe une corrélation entre les indices I_{CGFS} et I_{VMS}, on ajuste, dans chaque zone Z_i, le modèle linéaire sur ces indices, évalués aux points de l'échantillonnage CGFS. On ajuste ainsi par la méthode des moindres carrés le modèle linéaire

$$\mathbf{I}^{(\mathbf{i})}_{\mathbf{CGFS}} = a + b\mathbf{I}^{(\mathbf{i})}_{\mathbf{VMS}} + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma) \tag{3.39}$$

où la matrice Σ est la matrice de corrélation spatiale de l'indice I_{CGFS}. Cette matrice est donnée par un modèle de variogramme, ajusté (par les moindres carrés) à partir du variogramme empirique donné par l'indice I_{VMS}. Cette considération permet de prendre en compte l'autocorrélation spatiale présente dans les données, due à la faible distance entre certains points de l'échantillonnage. Notre intérêt est ici le signe du coefficient b dans la relation (3.39). Un paramètre b significativement positif indiquerait une corrélation linéaire positive entre l'indice d'abondance de la ressource et l'indice de perception de l'espace par les pêcheurs.

3.3.4 Résultats

Variogramme de l'indice I_{CGFS}

Le variogramme empirique (Petitgas, 1996) de l'indice I_{CGFS} a été ajusté sur les données. Une structure de variogramme exponentiel à effet de pépite a ensuite été ajustée par la méthode des moindre carrés. Ces deux variogrammes sont représentés sur la figure 3.13. On peut voir qu'il n'y a pas de structure de covariance spatiale nette sur l'indice I_{CGFS}.

Ajustement du modèle de mouvement

Le modèle de mouvement GaP a été ajusté pour les deux zones considérées, pour $K = 1$ jusqu'à $K = 7$ sous-zones. Dans chaque cas, le maximum de vraisemblance est estimé en utilisant la procédure Monte Carlo EM décrite dans la section 3.2. L'évolution de l'approximation

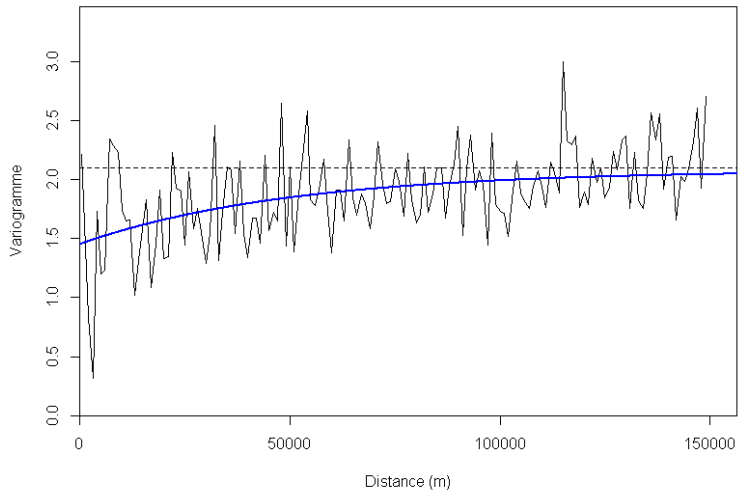


FIGURE *3.13 – Variogramme empirique et ajusté sur l'indice I_{CGFS}. La ligne brisée noire représente le variogramme empirique. La ligne bleue représente un variogramme exponentiel avec effet de pépite ajusté par moindre carrés.*

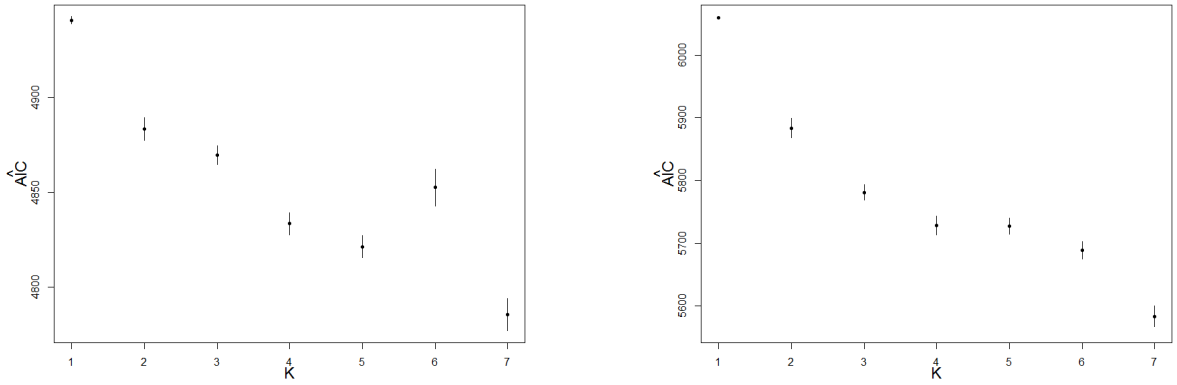
du critère AIC en fonction de K est montrée pour l'estimation sur les deux zones sur la figure 3.14. Dans les deux cas, c'est un modèle à 7 sous-zones qui est sélectionné.
Sur la figure 3.14a, on remarque une rupture dans la décroissance du critère AIC. Cette rupture est probablement due à un problème connu de l'algorihtme EM, qui est la convergence vers un maximum local de la vraisemblance. Dans ce cas, augmenter le nombres de points de départs de l'algorithme est conseillé [28].

De manière générale, pour cette méthode d'estimation du maximum de vraisemblance, il nous semble qu'il faut privilégier le nombre de points de départ par rapport au nombre d'itérations de l'algorithme MCEM. En effet, il semble que pour la majorité des points de départ, la convergence vers un maximum local est rapide. Cette heuristique nous semble d'autant plus importante que K (donc le nombre de paramètres à estimer) est grand. Un exemple de trajectoires de l'algorithme MCEM est montré en annexe E.2. Les figures 3.15b et 3.15d représentent les estimations des cartes de potentiel $P(\cdot, \theta_1)$ et $P(\cdot, \theta_2)$ guidant les pêcheurs ciblant la seiche pour la Baie de Seine et au large de Boulogne sur Mer, reconstruites à partir des trajectoires VMS (figures 3.15a et 3.15c).

Comparaison des indices I_{CGFS} et I_{VMS}

La figure 3.16 représente le vecteur $\mathbf{I}^{(\mathbf{i})}_{\mathbf{CGFS}}$ en fonction de $\mathbf{I}^{(\mathbf{i})}_{\mathbf{VMS}}$ pour chaque zone Z_i.
Le variogramme ajusté (figure 3.13) permet d'obtenir une matrice de corrélation Σ afin d'ajuster le modèle linéaire de l'équation (3.39). L'ajustement du modèle linéaire ne fait ressortir aucune corrélation significative entre les deux indices (paramètre b est estimé comme non significativement différent de 0).

28. Nous avons cependant laissé les résultats tels quels, afin d'illustrer ce phénomène



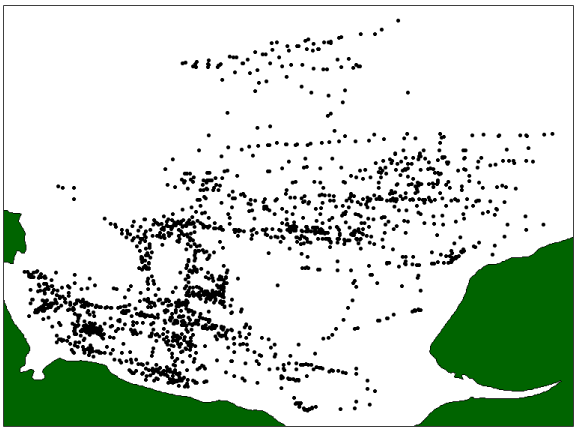
(a) Baie de Seine *(b) Boulogne sur Mer*

FIGURE *3.14 – Évolution de l'approximation du critère AIC en fonction de K pour l'ajustement du modèle GaP, sur les données de la Baie de Seine (gauche) et Boulogne sur Mer (droite). Le point représente la moyenne, les lignes représente le double de l'écart type du critère estimé. Dans les deux cas, c'est le modèle à 7 sous-zones qui est choisi.*

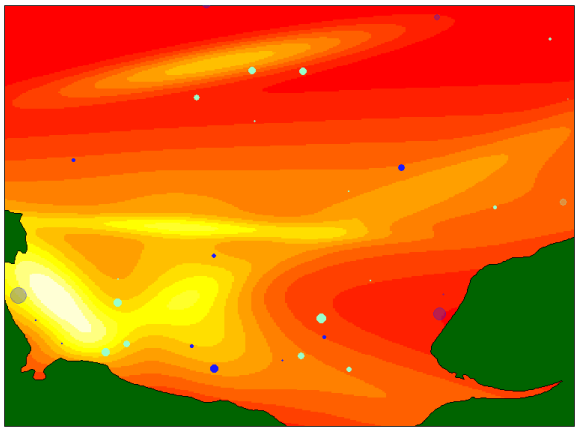
3.3.5 Discussion

Le travail présenté ici avait pour objectif de répondre à la question : La perception estimée de la distribution de la seiche en Manche Est au travers de l'activité des pêcheurs est elle corrélée à sa distribution estimée par la campagne scientifique ?
Si l'analyse n'a pas pu relever de corrélation significative, elle montre la pertinence méthodologique de l'approche du modèle GaP et tout son potentiel applicatif. En effet, le modèle GaP part de l'hypothèse que le mouvement est la conséquence d'une perception subjective de l'environnement. L'hypothèse faite par le modèle GaP est que le mouvement est guidé par le gradient de cette carte subjective, au travers du terme de dérive d'une EDS. En estimant cette perception subjective à partir des observations du déplacement, on peut donc la comparer directement à un indice estimé de façon indépendante à partir d'observations supposées être représentatives de la carte objective.

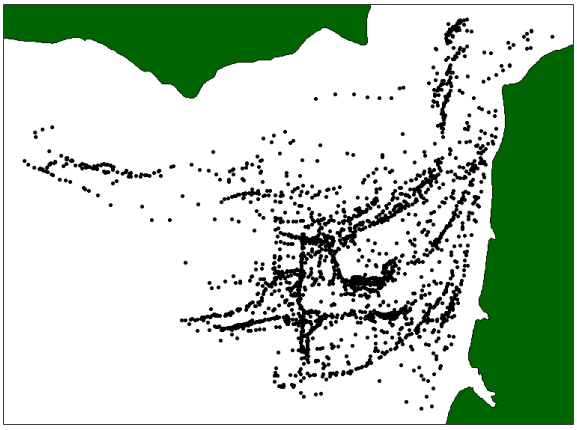
Contrairement aux méthodes classiquement utilisées pour estimer un champ d'activité de pêche, l'indice I_{VMS} de ce champ (au travers de la fonction P) est continu dans l'espace et ne nécessite aucune agrégation sur une grille (dont le choix de la taille de la maille est toujours sensible).
Cependant, l'hypothèse de formulation de P dans l'équation (3.24) induit des propriétés de symétrie de la carte de potentiel qui peuvent être limitantes pour expliciter la perception de l'environnement. De grandes valeurs de K rendent les formes des cartes plus flexibles (et donc moins symétriques) mais augmentent (linéairement) le nombre de paramètres à estimer. De plus, la forme en $\exp(-x^2)$ de la fonction P induit des cartes avec des propriétés fortes de régularité pouvant être incompatible avec la réalité des observations. Elles ne permettent notamment pas de représenter l'existence de discontinuités dans l'espace, qui pourraient être engendrées par des contraintes naturelles ou réglementaires. Cependant, comme précisé dans la section 3.2.5, un changement de forme fonctionnelle pour P n'est pas un changement très coûteux.



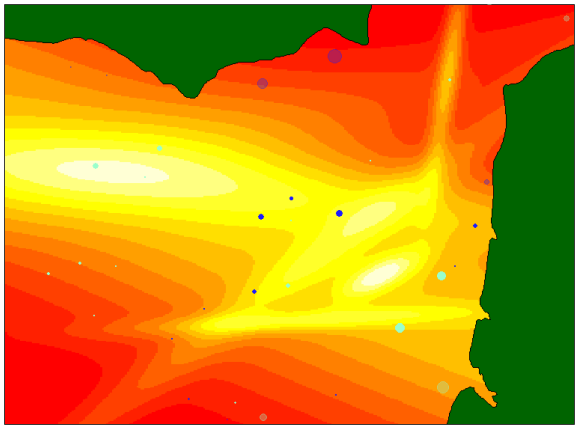
(a) Trajectoires observées (Baie de Seine)



(b) Indices I_{VMS} et I_{CGFS} (Baie de Seine)

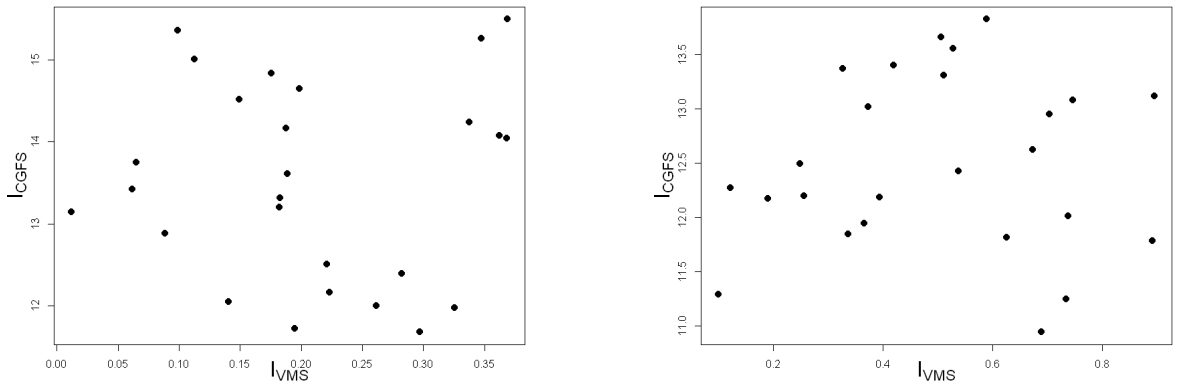


(c) Trajectoires observées (Large Boulogne)



(d) Indices I_{VMS} et I_{CGFS} (Large Boulogne)

FIGURE *3.15 – Trajectoires des navires de pêche ciblant la seiche Baie de Seine (resp. (a)) et au large de Boulogne (resp. (c)), et l'estimation du potentiel associé (resp. (b) et (d)). Un modèle à 7 sous-zones a été ajusté dans les deux cas. Les régions en rouge sont à faible potentiel, les régions en blanc sont à fort potentiel.*
Les points sur les cartes (b) et (d) représentent les valeurs relatives de l'indice I_{CGFS}. Les points bleus foncés(resp. clairs) sont représentent les valeurs au dessous (resp. au dessus) de la moyenne. La taille des cercles est d'autant plus grande que l'écart à la moyenne est grand. Les cercles transparents sont utilisés pour estimer la matrice Σ de l'équation (3.39) mais pas pour l'étude de la corrélation, car ils sont considérés comme inaccessibles aux navires de pêche. Les cercles pleins sont utilisés à la fois pour l'estimation de la matrice Σ et l'étude de la corrélation.



(a) Baie de Seine *(b) Boulogne sur Mer*

FIGURE *3.16 – Comparaison de l'indice I_{VMS} et de l'indice I_{CGFS} aux points d'échantillonnage de la campagne scientifique CGFS. À gauche, pour la Baie de Seine, à droite, pour Boulogne sur Mer. Aucun des deux cas ne révèle de corrélation significative.*

Un inconvénient majeur de cette méthode est sa forte paramétrisation. Ce problème est inhérent aux modèles mécanistes paramétriques. Dans notre cas cependant, le choix pour la fonctionnelle P induit une augmentation linéaire du nombre de paramètres avec K. L'estimation par maximum de vraisemblance peut alors devenir problématique dans la pratique. En effet, la maximisation dans un espace en grande dimension peut être difficile et coûteuse en temps. Pour un temps de calcul donné, on reste contraint par le nombre de points initiaux possibles pour l'algorithme, et il est alors possible que l'estimateur retenu corresponde à un maximum local, et non global, de la vraisemblance. Par exemple, dans l'étude présentée ici, pour le cas de la Baie de Seine, on observe un décrochement dans la monotonie de l'AIC pour $K = 6$ (figure 3.14a). Ce décrochement est potentiellement dû à l'atteinte d'un maximum local de la vraisemblance par la procédure MCEM. Il faudrait ainsi augmenter le temps d'application de la procédure pour augmenter les chances d'estimer le maximum global de la vraisemblance.

Cet exemple d'application a pu mettre en évidence l'apport de l'estimateur de l'AIC proposé à l'aide des équations (3.35) et (D.20). Dans l'exemple traité, il aurait peut être été judicieux de réaliser l'estimation en considérant un nombre de sous-zones plus important ($K > 7$), car aucun plateau ne semble avoir été atteint par le critère AIC. Cependant, cette augmentation pose ici la question du nombre de paramètres estimables pour une quantité fixe d'observations. Cette question mériterait sans doute une recherche méthodologique en perspective de celle réalisée en section 3.2

D'un point de vue pratique, nous n'avons pas mis en évidence de corrélation linéaire entre les les deux indices. En utilisant des méthodes différentes, mais avec des données comparables, (Joo, 2013) n'avait pas non plus pu expliciter de relation entre un champ spatial (proxy) d'abondance issu des données VMS, et un champ spatial d'abondance obtenu par campagne scientifique. Encore une fois, au vu du caractère restreint de l'étude, il n'est pas question ici de

donner de conclusions générales à la question posée [29]. Cependant, nous pouvons discuter ici de plusieurs points critiquables de notre approche.

Une première liste de limites concerne la manière dont la perception des pêcheurs a été estimée :

- La définition de I_{VMS} : Comme dit plus haut, la calcul de cet indice se base sur un modèle de mouvement. Or ce modèle de mouvement fait l'hypothèse que l'individu suit le gradient d'une carte lisse. Cette hypothèse est très forte, et amène une régularité dans les cartes estimées. Cette régularité a pour effet de ne pas induire de 0 systématique pour un endroit non visité. Cet effet, s'il est discutable en terme d'interprétation, a cependant l'avantage de faciliter les analyses de corrélation. Les méthodes à noyaux et de Ponts Browniens ont, en ce sens, tendance à procurer des cartes beaucoup plus "rugueuses", et pourraient être envisagées à titre comparatif.
- La non pénalisation des zones distantes : Dans cette étude, la notion de distance au port n'intervient pas, or les implications économiques d'un déplacement lointain peuvent pousser un individu à ne pas de déplacer sur une zone. Ainsi, il pourrait être envisageable de pondérer les indices par un fonction de coût pour prendre en compte cet effet d'augmentation des coûts d'exploitation par la distance.
- Le regroupement des navires : Dans cette étude, nous avons regroupé tous les navires ciblant la seiche, supposant leur mouvement issu d'un même modèle. Cette hypothèse ne prend ainsi pas en compte les disparités possibles entre navires. De plus, l'interaction entre les navires n'est pas considérée. Hors il est connu que certaines coopérations existent, et ont une influence sur l'activité (Laukkanen, 2003). Il serait intéressant d'étudier les manières d'intégrer dans cette étude de telles interactions, et d'en évaluer les impacts sur les résultats obtenus.

Une deuxième liste de limites concerne l'étude de la corrélation entre les deux perceptions estimées :

- Un problème de représentativité ? La couverture spatiale de l'échantillonnage CGFS est peut être ici insuffisante pour pouvoir être comparée avec l'activité de pêche. Ainsi, dans les exemples montrés ici, l'analyse des corrélations ne prend pas en compte des zones de forte concentration de l'activité de pêche, car ces zones ont une étendue spatiale qui n'est pas en adéquation avec l'échantillonnage CGFS proposé.
- Un indice en log : L'indice I_{CGFS} est calculé à partir d'une log transformation de la quantité de seiche par coup de chalut (standardisée par la surface chalutée). C'est donc un indice d'abondance dans échelle logarithmique. Cette transformation se justifie car elle permet de lisser des variations parfois grandes dans la valeur de l'indice avant transformation. Cependant, elle pose des problèmes d'interprétation car une variation linéaire de l'indice I_{CGFS} est à interpréter comme une variation linéaire de l'ordre de grandeur de l'abondance. Une perspective prioritaire de ce travail est de tester l'existence d'une

29. Comme beaucoup de travaux, celui ci apporte plus de questions que de réponses

relation entre l'indice I_{VMS} et l'indice présenté ici, mais non log-transformé.

– Champ moyen contre réalisation unique d'un champ aléatoire : La méthode proposée ici permet, à partir des trajectoires, d'estimer un champ de potentiel moyen pour le déplacement des navires. Ce champ est ensuite comparé à un indice d'abondance estimé en un point d'échantillonnage. Cette dernière valeur est la réalisation unique d'une variable aléatoire (la quantité de seiches pêchée), dont la variance est ici inconnue. Il est possible que cette variance soit trop forte, et empêche l'émergence d'un signal dans la comparaison entre les deux indices calculés ici.

Une perspective d'amélioration pour le dernier point serait d'intégrer un modèle de prédiction de l'abondance de la seiche conditionnellement aux observations scientifiques. Cependant, un tel modèle, de type krigeage (Krige, 1951), se baserait sur la structure de corrélation spatiale des données de seiche. Or, le variogramme montré en figure 3.13 montre que s'il existe une structure spatiale dans la répartition de la seiche, celle-ci n'est pas décelable à l'aide de l'échantillonnage proposé par la campagne CGFS. Il nous semble que cette impossibilité peut être due à deux facteurs : i) la structure de corrélation spatiale de la répartition de la seiche est à une échelle trop fine par rapport à l'échantillonnage CGFS ; ii) la variance dans la capture des seiches est telle, que tout signal de corrélation spatiale est indétectable.
Quelle que soit la raison sous-jacente, l'absence de structure spatiale semble empêcher ici la prédiction de l'abondance en des points extérieurs à l'échantillonnage, c'est pourquoi aucun modèle de krigeage n'a été adopté dans cette étude.

Ce travail permet de s'interroger l'hypothèse de distribution idéale libre pour les pêcheurs ciblant une ressource halieutique particulière. Les travaux corroborant cette hypothèse ont été effectués à des échelles spatiales larges (Swain and Wade, 2003). Dans ce travail, nous avons chercher à tester cette hypothèse à une échelle fine, rendue possible par les données VMS. Comme dans Joo (2013), aucune corrélation n'a pu être mise en évidence entre la distribution de la ressource et celle de l'activité de pêche. Il est possible qu'à cette échelle, des facteurs autres que la ressource impactent l'activité exercée par les pêcheurs, et que l'hypothèse de distribution idéale libre ne soit plus valable.
Mais les nombreuses limites de notre étude nous incitent à conclure par le traditionnel "further research is needed".

3.4 Discussion générale sur le lien entre champ spatial et trajectoire

Ce chapitre a été l'occasion de se pencher sur le développement d'un nouveau modèle de mouvement pour l'halieutique. Ce modèle, où une carte de perception est définie et estimée à partir des trajectoires, a été introduit à la suite des cartes d'utilisation. Ces cartes d'utilisation, résultant des trajectoires, ont un intérêt direct dans la connaissance de l'occupation de l'espace par un individu, à une échelle temporelle (supposée) globale, ou locale (sur le temps de l'observation). On obtient donc un champ spatial à partir des observations de la trajectoire.
Notre objectif de modélisation pour compléter ces méthodes partait de l'idée que les trajectoires sont la conséquence de l'existence d'un champ spatial qui les détermine. Dans le modèle proposé, nous formulons l'hypothèse que l'on peut définir une surface de potentiel qui dirige les trajectoires des individus, et nous avons appelé cette surface une carte de perception.
Le modèle pose ici plusieurs questions :

- **Lien avec le calcul de densité d'utilisation :** Les objectifs sont différents, mais les hypothèses sont-elles comparables ?
- **Hypothèses dues à la paramétrisation du modèle :** Quelles hypothèses le modèle GaP développé fait sur le mouvement ?

3.4.1 Densité d'utilisation et modèle de mouvement

La densité d'utilisation est liée à un modèle de mouvement

La densité d'utilisation de l'espace, qu'elle soit globale ou locale, est une résultante des déplacements de l'individu. À ce titre, elle est liée au mouvement. Des hypothèses sont ainsi faites sur la nature du mouvement, sous la forme d'un modèle parfois peu explicité.
Il nous semble important de préciser que l'estimation de la densité d'utilisation, même si elle ne s'intéresse pas directement à la modélisation du mouvement, n'échappe pas à des hypothèses fortes.
Dans le cadre de l'estimation non paramétrique par noyau de la densité d'utilisation globale, l'hypothèse est que les observations $X_0, \ldots, X_n$ aux temps $t_0 = 0, \ldots, t_n = T$ sont toutes des variables aléatoires de même densité p, indépendantes du temps. On fait donc l'hypothèse d'un modèle de mouvement satisfaisant cette condition, c'est-à-dire un modèle de mouvement strictement stationnaire [30].
D'un autre côté, pour calculer une densité d'utilisation locale, la méthode des ponts Browniens suppose que le mouvement observé est un mouvement Brownien. Ce modèle de mouvement

30. Cette hypothèse est très forte. Un exemple satisfaisant cette propriété est le processus de Ornstein Ulhenbeck, solution de l'EDS :

$$dX_t = \rho(\mu - X_t)dt + \sigma dW_t, \quad X_0 \sim \mathcal{N}(\mu, \frac{\sigma^2}{2\rho} I)$$

Le processus de Ornstein Ulhenbeck est asymptotiquement stationnaire si X_0 a une autre loi.

correspond à une équation différentielle stochastique "minimale", au sens des paramètres :

$$dX_t = \sigma dW_t.$$

L'hypothèse faite sur le mouvement est très forte mais les propriétés du mouvement Brownien font que sa densité d'utilisation est facilement calculable.
Le modèle GaP que nous avons présenté est un modèle de mouvement, au même titre que le mouvement Brownien. Ainsi, au même titre que la méthode des ponts Browniens, on pourrait calculer une densité d'utilisation d'un individu conditionnellement au modèle et aux observations. En effet, dans l'équation 3.13, la densité $p(z, \dots)$ pourrait être remplacée par celle définie sous le modèle GaP conditionnellement aux observations. Cette densité ne peut pas être exprimée de façon analytique mais pourrait être estimée par méthode de Monte Carlo [31], en utilisant les algorithmes de simulation conditionnelle exacte. On définirait ainsi une nouvelle fonction $h(z)$ qui serait une densité d'utilisation, celle obtenue par le pont du GaP. Cette démarche est ainsi possible pour n'importe quel modèle de mouvement.

Dans la littérature existante sur l'étude des "home range", la dépendance de la densité d'utilisation à un mouvement sous-jacent n'est pas toujours bien explicitée. Ainsi, il nous semble essentiel de souligner que le calcul de la densité d'utilisation est toujours sujette à un modèle de mouvement. Les hypothèses sous-jacentes de celui-ci sont parfois implicites (dans l'estimation non paramétrique) ou explicites (ponts Browniens).

Sur la stationnarité des processus

Les différentes approches présentées ici, méthodes par noyau, pont Brownien, ou modèle GaP, se distinguent justement par la nature de ces hypothèses sous-jacentes, notamment concernant la stationnarité du processus aléatoire de déplacement.
Un processus stochastique $(\mathcal{X}_t)_{t\geq 0}$ est dit strictement stationnaire si la distribution de la variable aléatoire $\mathcal{X}_t$ ne dépend pas du temps, et si la covariance du couple $(\mathcal{X}_t, \mathcal{X}_{t+\Delta})$ ne dépend que de Δ. On dit qu'il admet une distribution stationnaire si cette propriété est vraie quand t tend vers l'infini [32].
Dans la méthode de l'estimation par noyau, on fait l'hypothèse de stationnarité stricte. Ainsi, la position d'un individu dans l'espace a une densité de probabilité p qui ne dépend pas de l'instant d'observation. Cette hypothèse, très forte, permet de s'affranchir de la période d'observation pour calculer une densité d'utilisation moyenne et intemporelle.
En revanche, le mouvement Brownien n'admet de distribution stationnaire [33]. De ce fait on ne peut pas parler de densité d'utilisation de l'espace en faisant abstraction du temps. Néanmoins on peut toujours regarder la densité d'utilisation de l'espace conditionnellement aux observa-

31. L'intégration complète sur $[0, T]$ et l'espace serait probablement très difficile en pratique, pour des raisons de temps de calcul. L'important ici est la possibilité théorique de le faire
32. Plus précisément, si la loi de $\mathcal{X}_t$ converge vers la loi d'une variable aléatoire ne dépendant pas du temps
33. Uniquement une mesure stationnaire, la mesure de Lebesgue

tions, qui s'interprète comme l'utilisation de l'espace d'un individu sachant les lieux où il a été observé.
De la même manière, le processus solution du modèle GaP ne satisfait pas l'hypothèse de stationnarité. Ainsi, on ne peut pas utiliser le modèle GaP ou les ponts Browniens pour estimer une carte d'utilisation moyenne et intemporelle, mais seulement une carte d'utilisation conditionnelle aux observations.

Densité d'utilisation/carte de perception : Une confusion possible

Nous avons vu plus haut que le modèle GaP, par une méthode proche de celle des ponts Browniens, peut permettre la reconstruction d'une densité d'utilisation conditionnelle, donnant ainsi une carte d'utilisation classiquement utilisée en écologie. Il peut cependant exister une confusion entre une telle carte d'utilisation, et la carte de perception définie dans le modèle. Cette confusion peut notamment se manifester dans la volonté de comparer ces deux cartes. Ces deux cartes n'ont pas vocation à être comparées, car elles ne décrivent pas un même champ spatial.

Les densités d'utilisation sont à valeurs positives, et l'intégration de ces fonctions sur l'espace total donne la valeur 1. Elles décrivent intuitivement la distribution de présence d'un individu (conditionnellement aux observations, ou non, selon l'approche considérée). Les valeurs de la densité ont donc une interprétation "absolue" (le 0, en tous cas). Ces cartes ont une existence objective [34] au sens où les définitions de $\tilde{h}(z,T)$ et $h(z,T)$ sont toujours possibles pour une trajectoire continue.

Les cartes de perception de l'environnement sont, elles, issues de la conception du modélisateur. À ce titre, leur existence est conditionnelle au modèle, et philosophiquement plus discutable. Elle sont simplement définies comme des fonctions de potentiel (de $\mathbb{R}^2$ dans $\mathbb{R}$). Nous avons ici choisi de les présenter comme des fonctions bornées dans l'espace. Il convient de noter que cette carte de perception est définie, tel que nous l'avons fait, à une constante près. En effet, dans le modèle donné, les fonctions $P(x)$ et $P(x)+Constante$ donnent le même modèle de mouvement. La valeur d'une telle carte n'a pas une interprétation absolue, mais relative (relative aux valeurs voisines).

Le choix que nous avons fait pour la forme de la fonction P, dans les sections 3.2 et 3.3 est susceptible d'entretenir une confusion entre carte d'utilisation et carte des perceptions. Nous avons choisi une fonction positive dont la forme en $\exp\left\{(x-\mu)^T C(x-\mu)\right\}$, est proche de celle de la densité d'une loi Gaussienne. Ainsi, la confusion avec une densité de probabilité est possible (même si P n'intègre pas ici à 1 [35]). Comme discuté dans la section 3.2.5, le changement de forme pour P est une évolution mineure dans le modèle présenté ici (si P reste C^2 et bornée

34. Dans le paradigme stochastique
35. Elle n'est d'ailleurs même pas intégrable si on ne pose pas au moins $\lim_{||x||\to+\infty} P(x) = 0$

en x). À l'heure actuelle, P a des zones attractives et des zones neutres (auxquelles on peut donner la valeur arbitraire de 0). Un changement intéressant, qui écarterait toute confusion entre P est une densité de probabilité serait de proposer une forme avec zones attractives (à valeurs positives) *et* répulsives (à valeurs négatives). Cette forme pourrait intégrer des facteurs repoussant pour un individu. Par exemple, pour un chalutier en pêche, on peut imaginer qu'une zone de rochers joue un rôle répulsif, car c'est une zone inadaptée au chalutage.

3.4.2 Le modèle de mouvement GaP, quelles pertinence pour les hypothèses de déplacement ?

Une paramétrisation adaptée ?

Le modèle GaP se base sur une fonction de potentiel qui est un mélange de formes Gaussiennes. Ce modèle induit une croissance linéaire des paramètres avec le nombre de modes. Nous avons pu voir que les applications peuvent demander, pour bien décrire le mouvement observé, des surfaces potentiellement hautement multimodales. Ainsi, l'augmentation linéaire des paramètres pose un réel problème dans les cas d'applications, notamment concernant le temps de calcul nécessaire à l'estimation.
Dans le modèle proposé ici, une volonté était d'avoir des paramètres interprétables. Le modèle est régi par des paramètres de position (les μ), de forme (les matrices C) et de poids (les π). Il convient de noter que les deux derniers paramètres ont une influence sur le gradient de la surface de potentiel, et influencent donc la manière dont l'individu se dirige vers les centres attractifs. Ainsi, un individu aura tendance à se diriger vers les zones les plus attractives, ce qui était une hypothèse souhaitée, mais il aura aussi tendance à aller plus vite vers une zone très attractive (au poids π_k élevé) que vers une zone moyennement attractive (au poids π_k plus faible). Cette hypothèse induit un lien entre la vitesse d'un individu et l'attractivité des zones, ce qui ne correspond pas forcément à une hypothèse réaliste selon les cas d'application.

Un potentiel borné, une oasis dans le désert ?

Le cadre de modélisation estimation présenté ici est valable pour une fonction P de potentiel bornée.
En effet, dans le modèle présenté ici où le potentiel est un mélange de formes Gaussienne [36], un individu loin des zones attractives se retrouve dans une zone de gradient nulle. Ainsi, son mouvement loin des zones attractives est une marche aléatoire pure [37]. Loin d'une zone attractive, un individu erre donc comme dans un désert, incapable de rejoindre les oasis, si ce n'est par hasard. Ce problème advient pour tout potentiel continu borné, à moins que ce potentiel ne soit oscillant. Ce problème n'existe pas pour le processus de Ornstein Ulhenbeck par exemple (qui correspond à un potentiel non borné) où un individu, aussi loin soit-il du point attractif, aura

36. Ce terme est abusivement utilisé pour qualifier la forme de la fonction $\exp(-x^2)$
37. Quand le gradient est nul, le mouvement se résume au mouvement Brownien

tendance à s'y diriger [38]. Cette hypothèse est aussi une hypothèse forte, qui suppose un rayon d'attraction infini pour une zone attractive. Ce choix de modélisation par un potentiel borné est donc un ici un choix conceptuel fort, qui implique, en théorie, la possibilité d'un mouvement totalement diffusif des individus, hypothèse dont le réalisme est discutable.
Dans la pratique cependant, on peut considérer que l'individu étudié est toujours suffisamment proche de certains phénomènes d'intérêt, qui font que son mouvement ne sera jamais (dans la pratique) complètement diffusif.

Une perception uniquement locale ?

En définissant un mouvement guidé par un gradient, le modèle d'EDS basé sur un gradient de potentiel (borné ou non), implique que la perception qu'un individu a de l'environnement est purement locale. Les variations de l'attractivité de son environnement ne sont perçues par un individu que dans son voisinage direct. Et ce sont ces variations qui guident le mouvement. Le mouvement peut cependant se baser sur différents types de perception, chez un même individu. Lorsque le cycle de vie d'un individu est marqué par des mouvements à une échelle locale (phases sédentaires) et une échelle plus large (migration), la possibilité d'une perception uniquement locale est sans doute insuffisante.

Un potentiel homogène en temps ?

Dans le modèle décrit ici, on considère que le potentiel perçu par l'individu ne dépend pas du temps. Ainsi, le potentiel attribué à un point X de l'espace sera toujours identique. Cette homogénéité en temps pose problème pour décrire de manière réaliste des cycles marqués par des enchaînements de phases sédentaires et migratoires. Cette propriété pose aussi problème dans la description du mouvement des navires. En effet, dans un tel modèle, rien ne pousse un individu à retourner sur ces pas. Dans la pratique, comme sur les estimations montrées dans les sections 3.2 et 3.3, si ce temps de retour est faible comparé au temps total de la trajectoire, il semble que cela n'empêche pas une estimation de surface de potentiel interprétable.

Il n'en reste pas moins qu'un tel modèle ne pourrait pas être utilisé tel quel pour prédire des trajectoires de navire. En effet, le modèle ne décrit correctement qu'une partie du processus, celui de direction vers une zone et de son exploitation. On ne décrit ainsi pas le processus de retour vers la zone de départ (le port), qui semble incompatible avec l'hypothèse d'une surface de potentiel statique dans le temps. De plus, les mécanismes de transition entre zones sont ici techniquement possibles, mais ils peuvent être en pratique très long. En ce sens, il pourrait être intéressant d'induire un deuxième processus comportemental, qui permettrait le changement délibéré de zone. Cependant, ce processus devrait être formalisé en temps continu, ce qui complique sans doute la tâche par rapport aux modèles discrets de comportements pour l'activité de pêche. Une alternative serait d'intégrer la notion d'épuisement du potentiel d'une zone par une composante temporelle dans la fonction P, pour représenter par exemple l'épuisement local

38. Et d'autant plus vite qu'il en est loin

d'un patch de ressource. Cet épuisement entraînerait, de fait une tendance au changement de zone.

Sur l'ergodicité des processus

Un point important pour l'inférence des paramètres dans un modèle basé sur des EDS a été abordé dans la section 3.2.5, il s'agit de l'ergodicité des processus.
Un processus stochastique $(\mathcal{X}_t)_{t\geq 0}$ est dit ergodique si il existe une fonction de densité π telle que, pour toute fonction f,

$$\frac{1}{T}\int_0^T f(\mathcal{X}_t)dt \xrightarrow[T\to+\infty]{} \int_{\mathbb{R}^2} f(x)\pi(x)dx = \mathbb{E}_\pi(f(Z)) \tag{3.40}$$

où Z est une variable aléatoire de densité π. On appelle cette densité la loi stationnaire du processus. Si une telle densité existe, alors, la variable aléatoire $\mathcal{X}_t$ converge en loi vers Z. Un exemple de processus ergodique est le processus de Ornstein Ulhenbeck, où la variable aléatoire Z est alors celle d'une loi Gaussienne.

En revanche, le mouvement Brownien n'est pas ergodique. De même, le processus GaP n'est pas ergodique, ainsi qu'aucune solution d'une EDS

$$dX_t = \nabla P(X_t)dt + \gamma dW_t$$

où P est une fonction de potentiel bornée et C^2. Dans le cas du modèle GaP, une raison intuitive de la non ergodicité du processus est que la carte de potentiel induit, loin des zones attractives, des zones entièrement plates. En ces zones, le gradient de P est donc nul, ainsi le processus se comporte comme un mouvement Brownien, qui est non ergodique. Plus rigoureusement, si une distribution stationnaire π existait, elle serait proportionnelle à

$$\int_{\mathbb{R}^2} \exp\left\{\frac{P(x)}{\gamma}\right\} dx.$$

Or cette intégrale n'est pas définie si P est bornée. Donc π n'existe pas (Iacus, 2009).
Une conséquence de l'ergodicité est que l'observation du processus sur un temps long n'est pas suffisante pour en estimer toutes les caractéristiques. Ainsi, il est nécessaire d'observer plusieurs réalisations du même processus pour pouvoir garantir une bonne estimation des paramètres du modèle. En effet, si dans le cas ergodique, l'observation d'une trajectoire pendant un temps infini assure la convergence des estimateurs du maximum de vraisemblance, dans le cas non ergodique, la convergence est assurée quand le nombre de trajectoires observées tend vers l'infini.

Dans le cadre de l'application du modèle GaP présentée en section 3.3, nous disposions de plusieurs marées de plusieurs navires de pêche. En faisant l'hypothèse que chaque marée était une réalisation indépendante du processus solution du modèle GaP, on a pu alors disposer d'un large ensemble de trajectoires indépendantes pour l'estimation des paramètres. Il convient de

noter ici que pour la bonne estimation des paramètres du modèles GaP, de par la non ergodicité du processus solution, passe non pas par une observation longue d'une trajectoire, mais par l'observation de nombreuses trajectoires indépendantes.

3.4.3 L'approche continue par équations différentielles stochastiques, modélisation et estimation

Avantages d'une modélisation en temps continu

Le modèle développé est un processus en temps continu. On ne fait pas l'hypothèse que l'individu se déplace en unités de mouvements fondamentales (Getz and Saltz, 2008). L'approche proposée est inspirée par la notion de densité d'utilisation définie par l'équation (3.1). Cette approche tend à définir l'occupation de l'espace. Or si il peut apparaître *naturel* (McClintock *et al.*, 2014) que le comportement se déroule par phase [39] , il nous semble tout aussi naturel que l'occupation d'un lieu se fasse continûment dans le temps.
La modélisation continue permet ici une formulation du modèle indépendante du processus d'échantillonnage. La méthode d'acquisition des données n'intervient ainsi pas à travers la définition d'un pas de temps minimal, comme cela est nécessaire dans un modèle en temps discret. De la même manière, la méthode d'inférence est indépendante du pas de temps d'acquisition, [40] et ne requiert ainsi pas d'interpolation des données.
Cette approche continue, et la volonté de décrire les mécanismes du déplacement d'un individu, nous ont amenés à la formalisation d'un modèle à l'aide des équations différentielles stochastiques.
L'approche par EDS découle naturellement de la notion de marche aléatoire, présente depuis longtemps en écologie (Skellam, 1951). Elle permet une modélisation flexible de différents types de mouvement (Harris and Blackwell, 2013). En considérant notamment que la dérive de l'EDS est une surface de potentiel, le mouvement découle alors d'une fonction scalaire interprétable (Brillinger, 2010). C'est cette idée qui a été reprise, afin de quantifier la perception de l'environnement d'un individu. La manière dont nous l'avons modélisée entraîne cependant différentes conséquences. Une conséquence importante est que la solution d'une EDS, lorsqu'elle existe, est un processus Markovien. Cette contrainte rend difficile l'inclusion directe dans le modèle d'une mémoire à long terme chez un individu.

La méthode statistique d'estimation

Les méthodes d'estimation développées dans ce travail se distinguent des techniques utilisées dans la littérature dédiée à l'écologie du mouvement. Pour les modèles présentés ici, Brillinger (2010) et (Preisler *et al.*, 2013) utilisait des schémas d'Euler pour approcher les lois de transition du processus. Cette approximation était faite pour n'importe quel pas de temps d'acquisition, rendant difficilement quantifiable l'erreur d'approximation de la vraisemblance totale.

39. Justifiant l'idée de pas de temps fondamental
40. Dans sa formulation, sa performance peut évidemment être impactée

Dans le travail proposé ici, l'estimation des paramètres est faite par maximum de vraisemblance, en utilisant une approche EM couplée aux algorithmes de simulation conditionnelle exacte proposés par Beskos *et al.* (2006b). Cette procédure garantit une estimation sans biais du maximum de vraisemblance, si il est atteint. L'erreur peut alors provenir de 3 sources

1. Un mauvais choix de point de départ pour l'algorithme EM, qui mène à un maximum local de la vraisemblance ;
2. Une non convergence dans l'étape M. Celle ci étant effectuée par l'algorithme CMAES, elle revêt un caractère stochastique qui ne garantit pas, dans les cas complexes, l'atteinte du maximum ;
3. L'erreur d'approximation due à l'approximation de Monte Carlo ;

Concrètement, nous proposons 3 heuristiques pour contourner ces problèmes :

1. Multiplier le nombre de points de départ [41] ;
2. Multiplier le nombre de particules simulées dans l'algorithme CMA-ES, ou, dans certains cas, calculer le gradient de la fonction à maximiser [42] ;
3. Augmenter le nombres de particules simulées.

Dans la pratique, et de manière purement heuristique, il nous a semblé que la première étape était cruciale pour une application aux données réelles, a fortiori quand le nombre de modes est grand.
La durée de l'étape M est essentiellement dépendante du nombre de modes et de particules simulées. La durée de l'étape E est extrêmement variable. Concrètement la simulation conditionnelle sera d'autant plus longue que

1. Le relief du potentiel P est marqué. Dans la paramétrisation proposée ici, cela se traduit par des grandes valeurs pour les $(\pi_k)_{k=1,\dots,K}$ et/ou des grandes valeurs de déterminant des $(C_k)_{k=1,\dots,K}$.
2. La diffusion est faible par rapport à la dérive. Cela se traduit par un γ faible.

Intuitivement ces deux conditions traduisent un écart au mouvement Brownien. Le mouvement Brownien peut se voir comme la solution de la même équation que le modèle GaP, mais avec une carte de potentiel complètement plate, et, par conséquent, une diffusion (infiniment) plus dominante que la dérive. Or , l'étape E se base sur un algorithme d'acceptation rejet dont la loi de proposition est celle d'un mouvement Brownien [43]. Il n'est donc pas surprenant que cette étape soit plus longue quand la loi cible est très éloignée de celle du Brownien.
Ainsi, pour certains sous ensembles de paramètres, cette étape peut être très couteuse. Dans cette optique, il serait peut être judicieux d'utiliser l'algorithme SAEM, qui nécessite la simulation de moins de particules que l'algorithme MCEM (Delyon *et al.*, 1999). Cet algorithme, couplé à un schéma d'Euler a déjà été proposé pour l'estimation de paramètres dans des modèles d'EDS pour la biologie (Ditlevsen and Samson, 2014). Cette optimisation de la performance en

41. Et les choisir intelligemment !
42. Ce gradient peut avoir une forme effrayante
43. Conditionné par l'arrivée, mais mouvement Brownien tout de même

temps pour l'estimation apparaît essentielle si on souhaite pouvoir appliquer cette technique de manière opérationnelle sur de grands jeux de données.

3.4.4 Bilan : Le modèle GaP, une base intéressante pour la modélisation-estimation en temps continu

Pour conclure, il nous semble que le modèle présenté constitue une base intéressante pour modéliser le mouvement des individus avec un mécanisme simple [44]. Ce modèle en temps continu s'appuie sur une hypothèse réaliste, est construit à partir de paramètres interprétables et est formulé indépendamment de la méthode d'échantillonnage. De par sa flexibilité et sa généralité, nous pensons que le formalisme des EDS est adapté à la description du mouvement individuel. Sa formalisation permet d'envisager des perspectives de développement très riches, qui seront discutées dans la section 4.2.

Dans ce travail, un effort important a été réalisé pour développer une méthode d'inférence statistique pour ce modèle mécaniste de déplacement. Ce chapitre démontre la faisabilité de l'inférence pour un tel modèle, ce qui participe largement de son intérêt pour stimuler de nombreuses applications à de nombreux cas d'études.

Le développement des méthodes d'inférence, qui mobilise le cadre des modèles à espace d'états, montre de grande similarités avec le cadre utilisé pour l'inférence des HMM, plus connu et plus développé dans la littérature statistique et appliquée à l'écologie. La mise en évidence de ces similarités rend le problème d'estimation des EDS, de notre avis, abordable pour les utilisateurs. En ce sens, nous espérons que l'aspect technique des modèles en temps continu qui pouvait rebuter les utilisateurs (McClintock *et al.*, 2014), est désormais moins effrayant [45].

44. Simpliste ?
45. En tous cas, il l'est moins pour l'auteur

Chapitre 4

Les modèles de mouvement, perspectives d'utilisation et d'amélioration

Ce travail de thèse s'est intéressé à la formulation de modèles stochastiques et mécanistes pour l'analyse de trajectoires, en s'appuyant sur des applications halieutiques. Les limites des modèles développés dans les chapitres 2 et 3 ont été exposées dans les sections 2.4 et 3.4. Sans y revenir en détails, nous pouvons ici insister sur un caractère essentiel des modèles de mouvement abordés ici.

L'ambition de la démarche (ou des démarches) de modélisation est double : il s'agit d'abord de décrire les mécanismes du mouvement au travers d'une formulation paramétrique des causes du déplacement. Mais les modèles proposés doivent également être compatibles avec une démarche d'inférence statistique. Les paramètres régissant le mouvement doivent alors être estimés à partir des observations des trajectoires.

Ainsi, les modèles développés dans cette thèse n'échappent pas à la nécessité de réaliser des compromis entre réalisme du modèle (plus de réalisme impliquant souvent plus de complexité et de paramètres) et faisabilité de la démarche d'inférence.
La formulation des modèles de mouvement s'est donc appuyée sur des hypothèses qui traduisent une simplification de la réalité. Cette simplification a deux origines non nécessairement distinctes :

L'adéquation entre la formulation du modèle et la question posée : La formulation d'un modèle a pour but une réponse à une ou plusieurs questions. Dans notre cas d'étude, nous pouvons distinguer deux questions principales

1. La question de la segmentation des séquences des comportement. Comment détecter les moments d'activité de pêche au cours d'une trajectoire ?
2. La question de l'inférence d'un champ de potentiel. Comment déterminer la manière dont l'espace est perçu par les l'individu ?

Ainsi, la modèle formulé étant centré sur la question posée, il peut mal spécifier (ou ne pas les spécifier du tout) des phénomènes supposés non centraux relativement à cette question. Un exemple de ce cas de figure est discuté dans la section 2.4.3, où la durée des temps de séjour d'un navire dans chacun de ses comportements est mal spécifiée. Cependant, il est intéressant de noter qu'une meilleure spécification de ces lois par un modèle de semi Markov n'améliore pas la capacité de détection de l'activité de pêche (figures 2.16 et 2.17).

L'estimabilité des paramètres à partir des données : Dans ce travail, la capacité à réaliser des inférences sur les paramètres à partir d'observations de trajectoires est posée comme un objectif fondamental. Ainsi, le modèle proposé doit pouvoir être estimé, en théorie, comme en pratique. Dans les exemples traités ici, et de manière générale, la seule donnée récurrente consiste en une ou plusieurs séquences de positions observées, sur un pas de temps discret, régulier ou non.
L'objectif est alors de formuler un modèle dont les paramètres sont estimables à partir de ces seules séquences de positions. En effet, le cadre d'inférence étant ici non Bayesien, aucune information a priori n'est intégrée sur les paramètres. Cette condition impose de poser un modèle identifiable en théorie, au sens où deux ensembles de paramètres différents θ_1 et θ_2 ne doivent pas donner la même loi du mouvement, c'est-à-dire la loi du processus $\mathcal{X}$. À ce besoin d'identifiabilité théorique, s'ajoute un besoin d'identifiabilité pratique. Le modèle spécifié doit pouvoir être estimé à partir des données effectivement disponibles.

Cette simplification de la réalité est donc à la fois un choix et une nécessité. Nous proposons ici des pistes de perspectives pour ces modèles, qui permettraient de lever certaines hypothèses restrictives, et/ou d'enrichir les processus de nouveaux mécanismes. À ces perspectives conceptuelles sur la définition de nouveaux modèles, nous ajoutons des perspectives opérationnelles pour les modèles que nous avons développés. Ainsi, nous essayons de situer nos travaux par rapport aux outils existants, et leur potentiel pour une utilisation directe en halieutique, et plus largement en écologie du mouvement. Ces perspectives sont développées selon deux axes différents, les modèles de détection de l'activité, et les modèles de reconstruction d'un champ spatial.

Les chapitres précédents ont montré que l'intérêt essentiel des modèles HMM, en halieutique et en écologie, porte sur la modélisation de la séquence des différents comportements régissant une trajectoire. Ces modèles mettent généralement l'accent sur la dispersion temporelle, la dimension spatiale devenant souvent accessoire. D'un autre côté, le modèle de mouvement dérivant d'une carte de potentiel met l'accent sur la modélisation des mécanismes spatiaux régissant le mouvement, mais reste pauvre en ce qui concerne la modélisation des séquences de comportements différents le long de la trajectoire.
Ainsi, c'est assez naturellement que les perspectives ouvertes sur le cadre HMM (développées dans la section 4.1) portent sur la modélisation des transitions entre comportements, tandis que les perspectives ouvertes par le modèle GaP portent plus directement sur la fonction de

potentiel gouvernant le mouvement (section 4.2). Dans la section 4.3, nous traçons des pistes de réunion de ces deux approches, qui permettraient d'intégrer un modèle de transition entre comportements à un modèle spatialisé du déplacement. Enfin, nous conclurons ce manuscrit dans la section 4.4.

4.1 Modéliser et reconstruire les comportements

4.1.1 Perspectives opérationnelles pour l'halieutique : Les modèles HMM, la meilleure des méthodes pour détecter l'activité de pêche ?

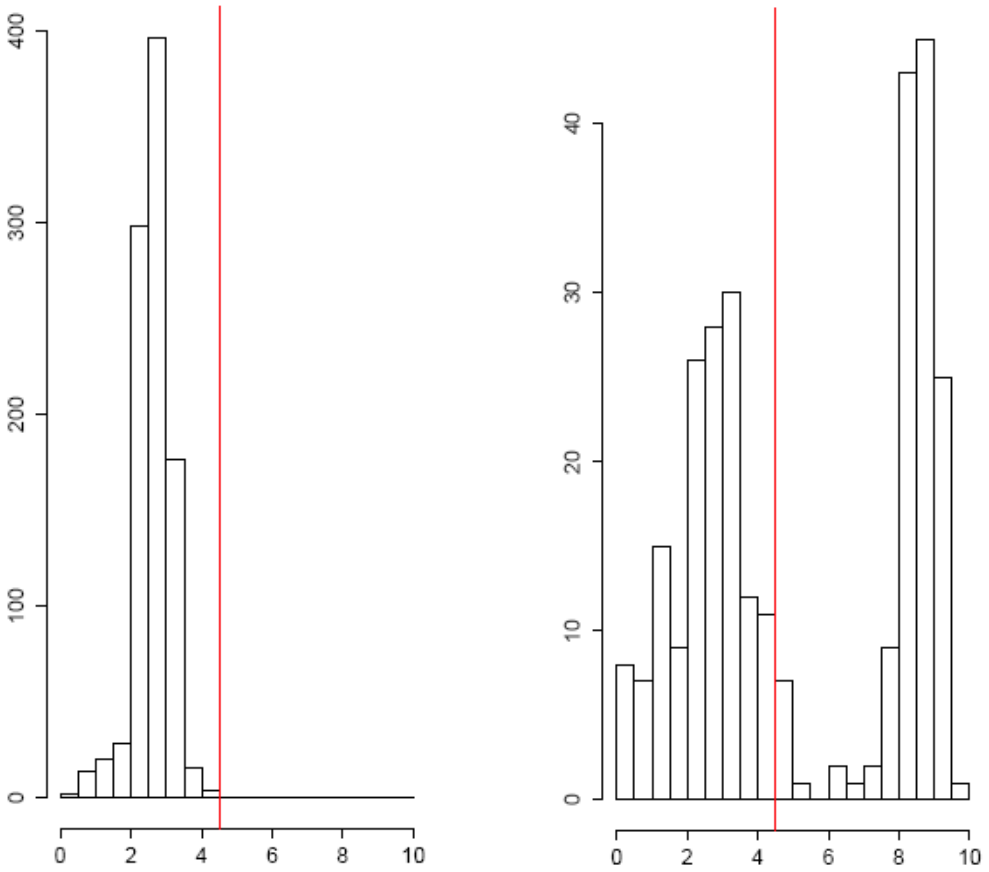
Une segmentation dichotomique, le modèle de seuil

Le chapitre 2 traite du développement d'un modèle HMM pour détecter l'activité de pêche des navires en Manche-Est. Deux principales alternatives ont été proposées dans la littérature pour détecter l'activité de pêche à partir des données VMS :

1. Modèle de seuil : Cette méthode consiste à appliquer un seuil sur les vitesses du navire (Palmer and Wigley, 2009). Si la vitesse (le plus souvent calculée) en un point est en deça d'un certain seuil, alors le navire sera considéré en pêche pour le point donné. Cette méthode, peu gourmande en calcul, est la plus utilisée, et sert aujourd'hui pour le traitement opérationnel des données VMS à grande échelle (dont l'échelle nationale pour la France, Berthou *et al.* (2013)) ;
2. Classification supervisée par réseaux de neurones ; À l'aide d'une base de données d'apprentissage, la position d'un navire est dite en pêche ou non, en fonction de caractéristiques telles que sa vitesse, sa direction, son accélération (en général, ces quantités sont calculées, à partir des séquences de positions observées). Cette méthode a par exemple été proposée dans le cadre de la pêcherie de l'anchois (Bertrand *et al.*, 2008; Joo *et al.*, 2011)

Les résultats de la littérature montrent que les performances de ces méthodes sont très dépendantes de la pêcherie considérée. La méthode des réseaux de neurones a montré des résultats bien meilleurs que la méthode de seuil dans le cadre de la pêcherie d'anchois du Pérou. Une des raisons est l'existence d'une phase de recherche des bancs de poissons associée à la pêche à l'anchois. Cette phase induit une vitesse faible des navires, qui était souvent considérée comme une phase de pêche par la méthode de seuil. Un inconvénient majeur de cette méthode est qu'elle demande une large banque de données d'apprentissage.

Dans le cadre de la pêcherie de fond de la Manche Est, aucune base de données suffisante pour la classification supervisée n'est disponible. Pour notre zone d'étude, aucune analyse complète utilisant une large base de données de validation, et qui permettrait de connaître les performances de la méthode de seuil, n'a été réalisée à ce jour.
Cependant, de notre expérience sur un ensemble restreint de données, la méthode de seuil pro-



(a) Pêche *(b) Non pêche*

FIGURE *4.1 – Histogrammes représentant la distribution globale des vitesses d'un chalutier en Manche Est lorsqu'il pêche (à gauche), et lorsqu'il ne pêche pas (à droite). La ligne verticale est la vitesse (4.5 noeuds) en deçà de laquelle il est estimé "en pêche" par le traitement générique des données VMS (Berthou* et al.*, 2013). On surestime donc l'activité de pêche. L'axe des ordonnées représente le nombre de positions observées par classe. Notez la différence d'échelles entre les deux graphiques. La connaissance de l'activité est procurée par un observateur embarqué du programme Obsmer (Dubé* et al.*, 2012)*

cure une erreur de classification de l'ordre de 10 à 15% pour notre cas d'étude (le tableau 2.4 montre les résultats pour une trajectoire). Un autre résultat important est que la méthode de seuil a tendance à biaiser les estimations dans le sens d'une surestimation de l'activité de pêche. En effet, le seuil de vitesse adopté pour cette pêcherie est de 4.5 nœuds. Aucun navire en pêche n'a une vitesse excédant 4.5 nœuds. D'un autre côté, il peut arriver qu'un navire ne soit pas en pêche, malgré une vitesse en deçà de ce seuil. La figure 4.1 montre la distribution globale des vitesses d'un chalutier dans la Manche Est selon son activité (obtenue grâce à un observateur embarqué du programme Obsmer (Dubé *et al.*, 2012)), et l'application de la méthode de seuil à ce navire.

Dans un cadre de dichotomie "pêche", "non pêche", le modèle de Vermard *et al.* (2010) peut être essentiellement considéré comme une extension stochastique de la méthode de seuil. De notre expérience, restreinte encore une fois, cette méthode a des performances similaires à la méthode de seuil sur les données de Manche Est. Quant au modèle AR HMM, développé dans la

section 2.2, ses performances ont été discutées dans les sections 2.3 et 2.4, et sont généralement moins bonnes que les deux méthodes précédemment citées. Ceci étant, pour généraliser les résultats obtenus sur une base de données restreinte de trajectoires, une étude étendue mériterait d'être réalisée.

À la lumière des résultats obtenus ici il semble que d'un point de vue opérationnel, la méthode des seuils des vitesses procure, pour les chalutiers de fond de la Manche Est, une qualité de détection de l'activité de pêche satisfaisante, qui n'est pas améliorée par les méthodes HMM. Or, contrairement au cadre HMM, la méthode de seuil ne nécessite aucun développement spécifique en termes d'inférence, et reste donc d'un point de vue opérationnel, beaucoup plus rapide à mettre en œuvre. Ainsi, au vu du développement actuel des HMM, la méthode de seuil apparaît encore comme la meilleure méthode opérationnelle dans l'objectif de détecter l'activité de pêche à partir de grandes bases de données de trajectoires.

Au delà d'une simple dichotomie pêche/non pêche ?

Si l'approche HMM développée dans cette thèse s'est intéressée à la segementation de l'activité de pêche sous la forme de deux comportements différents pêche et non pêche, elle peut aisément être étendue pour distinguer un nombre supérieur de comportements. Pour répondre à la diversité des activités de pêche rencontrées, différentes variantes des HMM basées sur 3 comportements ont ainsi été développées en introduisant des états supplémentaires tels qu'un état de recherche, ou un état de transition (arrêt, entrée ou sortie d'un port).
Ces développements ont été proposés pour la pêcherie pélagique [1] du Golfe de Gascogne (Vermard *et al.*, 2010), les thoniers seineurs de l'océan Indien (Walker and Bez, 2010; Bez *et al.*, 2011; Walker *et al.*, 2015), la pêcherie de crevettes en Australie (Peel and Good, 2011), la pêcherie d'anchois du Pérou (Joo *et al.*, 2013a), et plus récemment, sur des engins fixes pour la pêche du crabe des neiges (Charles *et al.*, 2014).

Pour la pêcherie des chalutiers de fond de la Manche Est, on peut s'interroger sur quel comportement ou état ajouter pour sortir d'une simple dichotomie pêche/non pêche :

- Comportement de recherche ? La pêcherie de fond semble peu adaptée à un tel état, car il n'y pas de recherche (visuelle, ou par radar) de la ressource. Ainsi, la route vers la zone de pêche est bien plus directe que dans les pêcheries où la ressource est détectable depuis la surface (telle que la pêcherie d'anchois, ou thonière).
- Etat de transition (arrêt, court trajet entre deux zones) ? Cet état de transition pourrait permettre de réduire les faux positifs qu'induit la méthode de seuil. Cependant, le caractère ponctuel de cet état (durée par nature très courte) le rendrait sans doute très difficilement détectable au pas de temps considéré d'une heure ;
- Différents états de pêche ? Les pêcheries en Manche Est sont des pêcheries mixtes, et il

1. Un chalut pélagique est un chalut qui traîne dans la colonne d'eau, et non sur le fond. Il cible les espèces pélagiques (vivant dans la colonne d'eau). Dans le cadre de l'article de Vermard *et al.* (2010), il s'agit d'anchois, de sardines, de thons, ou de maquereaux

arrive que les chalutiers de fond utilisent 2 engins différents. Il arrive également qu'un chalutier cible différentes espèces au cours d'une même marée. D'un point de vue opérationnel, on peut se demander si l'ajout d'un troisième comportement pourrait permettre de détecter des différences dans les activités de pêche à deux niveaux

1. Au niveau des engins ? Le navire change-t-il d'engin durant sa marée ?
2. Au niveau des espèces cibles ? Le navire change-t-il d'espèce cible au cours d'une marée ?

Ces questions spécifiques à la Manche Est ont été explorées au travers d'une collaboration réalisée dans le cadre du projet de recherche Européen VECTORS [2] (travail de Mathieu Woillez *et al.* dans Marchal *et al.*, 2014). Les résultats obtenus semblent montrer une capacité à distinguer deux types d'activité de pêche, en fonction de deux engins différents. En revanche, la capacité à distinguer des activités différentes en fonction des espèces cibles n'a pas pu être démontrée. De manière générale, cette modélisation plus fine des comportements de pêche, en interaction avec les experts de la dynamique des flottilles apparaît comme une perspective intéressante pour les approches HMM.

4.1.2 Perspectives pour les HMM en halieutique et en écologie

Vers des modèles prédicitifs de séquences de comportement ?

Une fois leurs paramètres estimés, les modèles HMM pour détecter l'activité permettent la simulation de trajectoires conjointement à la succession de différents types de comportements. La possibilité de simuler ces processus peut permettre l'étude de différents scénarios affectant le mouvement. Plusieurs perspectives de développement des HMM permettraient de proposer des cadres de simulation de séquence de comportements. Ces amélioration pourrait être intégrées dans un modèle prédictif global, qui comprendrait également un modèle de mouvement conditionnel à l'état (ce dernier point sera abordé en section 4.3). Nous proposons ici quelques perspectives pour l'amélioration de la modélisation des comportements.

Les lois de temps de séjour

Comme discuté dans la section 2.4, les modèles Markoviens développés induisent une loi de temps de séjour exponentielle dans chaque comportement. Si cette hypothèse ne semble pas être déterminante dans notre cas pour la bonne classification des états de pêche, elle demeure très restrictive pour la formulation d'un modèle prédictif pour la séquence de comportements. Les méthodes de semi Markov Caché semblent plus flexibles, afin de modéliser les temps de séjour dans les comportements par des lois plus conforme aux observations. L'estimation des paramètres du mouvement, et des lois de temps de séjour peut alors être faite par des méthodes similaires à celles présentées ici. Dans un formalisme en temps discret, les méthodes d'estimation de tels modèles ont été explicitées (Guédon, 2003). Plus récemment, Langrock and Zucchini (2011) ont montré que tout modèle semi-Markovien caché (HSMM) pouvait être approché par

2. `http://www.marine-vectors.eu/`

un modèle Markovien caché (HMM). En utilisant cette approximation, les méthodes d'inférence pour les HMM, maintenant bien établies en écologie (et utilisées ici dans le chapitre 2) restent donc applicables.

Les mécanismes de transition des comportements
Dans les modèles présentés ici, les mécanismes de transition entre comportements sont supposés être régis par une matrice Π dont les paramètres sont homogènes dans le temps et l'espace. Dans l'optique d'un modèle prédictif, il serait intéressant de modéliser les transitions par une matrice de transition inhomogène, dépendant de l'espace et/ou du temps.

a) Inhomogénéité spatiale des mécanismes de transition Une extension possible des modèles HMM est de faire dépendre la matrice de transition Π de la position dans l'espace x et de covariables $\lambda(x)$ évaluées en x. La matrice serait donc de la forme $\Pi(x, \lambda(x))$.

Exemple 1 Soit $\lambda(x)$ une fonction connue représentant une covariable mesurée en tout point x. Il est possible d'inclure la fonction $\lambda(x)$ dans la matrice Π typiquement grâce à une fonction logistique comprise entre 0 et 1. La probabilité de passer d'un état à un autre dépend alors de l'espace au travers de cette fonction. Par exemple, l'absence complète de ressource conduirait un prédateur à persister dans un état de route vers une zone de nourriture.
Pour une fonction logistique de la forme $(1 + a \exp(-r\lambda(x))^{-1}$, la matrice dépendrait donc des paramètres a et r. On peut supposer ces paramètres connus, ou inconnus et à estimer à l'aide des données disponibles.
Des modèles de ce type ont été proposés pour décrire le mouvement de cerfs (Morales *et al.*, 2004), et de thons (Patterson *et al.*, 2009). Les paramètres sont estimés à partir des observations grâce à un modèle hiérarchique Bayésien.

Exemple 2 Une modélisation permettant l'identification de points d'attraction de l'espace inconnu pourrait également être proposée. Soit μ_0 une position de l'espace riche en ressource, inconnue (donc à estimer). On peut par exemple supposer chez l'individu un comportement de route qu'on ne quitte qu'à proximité de μ_0, zone autour de laquelle l'activité d'exploitation s'intensifie. La probabilité de rester en route pourrait alors prendre la forme paramétrique $1 - \exp(- \parallel x - \mu_0 \parallel^2)$. Ainsi, le passage d'un état à l'autre serait informatif sur la proximité d'un point d'intérêt pour l'individu. Si on suppose que μ_0 influe aussi sur le mouvement conditionnellement à l'état (comme dans le modèle GaP, par exemple), les deux informations (transitions de comportements, et mouvement) pourrait alors concourir pour estimer le centre d'intérêt μ_0. Nous ne connaissons pas de tel modèle dans la littérature en écologie du mouvement.

b) L'inhomogénéité temporelle des mécanismes de transition En règle générale, en écologie du mouvement, l'hypothèse de mécanismes de transition entre les comportements homogènes dans le temps demande à être relaxée. Pour un navire de pêche, par exemple, la

probabilité de rentrer au port tend vers 1 quand le temps avance (par atteinte des limites d'autonomie du navire, cargaison pleine, etc...). Encore une fois, l'intégration d'une fonction logistique du temps $\lambda(t)$ dans la matrice de transition peut permettre d'intégrer cette non homogénéité. L'utilisation de modèles de Markov caché non homogènes a déjà été proposée, et des méthodes d'estimation proches de celles présentées ici ont été développées, dans un cadre de maximum de vraisemblance (Hughes *et al.*, 1999), ou dans un cadre Bayesien (Pinto and Spezia, 2015).

Des comportements associées à d'autres caractéristiques de l'état de l'individu ?

Les modèles HMM présentés ici consistent en une séquence de comportement que l'on considère comme guidant un processus $\mathcal{X}$, qui est la trajectoire de l'individu. Ceci étant, les comportements peuvent être associés à d'autres caractéristiques de l'individu que sa manière de se mouvoir. Il est alors envisageable d'adapter ces modèles tels que le processus $\mathcal{X}$ observé ne consiste plus seulement en les positions d'un individu, mais en un ensemble constitué des positions et d'un ensemble de caractéristiques de l'état d'un individu. En écologie, cette possibilité est attrayante du fait de l'émergence de bio-loggers ou de données biotélémétriques, permettant d'enregistrer un grand nombre de traits pour un animal (Cooke *et al.*, 2004). Ainsi, on peut mesurer l'activité de la mâchoire chez un crabe, la consommation d'oxygène chez un saumon, etc... Ces traits sont eux mêmes révélateurs du comportement. Il est donc naturel de les intégrer au déplacement pour enrichir les informations permettant de retracer le comportement d'un individu. Dans la pratique, les mesures de physiologie ne sont pas nécessairement synchrones à celles des positions. La formulation en temps discret peut donc s'avérer inadaptée dans ce cas. La modélisation en temps continu de tout les processus (comportements, mouvement, physiologie) est alors sans doute souhaitable pour un couplage en adéquation avec les réalités de l'échantillonnage.

4.2 Modéliser et reconstruire un potentiel

Le chapitre 3 traite du développement d'un modèle de déplacement guidé par un potentiel, similaire à ceux proposés par Preisler *et al.* (2013) pour l'écologie, et de l'application de ce modèle à l'halieutique.
Grâce aux travaux de Beskos *et al.* (2006b), nous avons pu développer une méthode d'inférence statistique robuste, qui ne fait pas d'approximation Gaussienne dépendante de l'échantillonnage. La pertinence de ce modèle a été discutée à la fin du chapitre 3. Ici, nous traçons quelques unes des nombreuses perspectives ouvertes par ce travail.

4.2.1 Perspectives opérationnelles pour l'halieutique : Les données VMS comme proxy de l'abondance ?

Dans la section 3.3, nous avons utilisé le modèle GaP pour déterminer si les cartes de perception des navires de pêche, estimées à partir des données VMS, donnait une information

sur la distribution de la ressource, estimée par campagne scientifique. Dans cette section, aucun lien n'a pu être démontré entre les deux approches. L'approche adoptée ici fait de nombreuses hypothèses dont il convient d'évaluer les impacts sur les résultats. À large échelle, il a été montré que l'effort de pêche était un bon proxy de l'abondance, pour certaines pêcheries (Swain and Wade, 2003). Notre étude, comme celle de (Joo, 2013, Chapitre 7) n'a pas mis en évidence, dans des contextes différents, de telle relation. Il convient de se demander si les résultats sont influencés par le modèle de mouvement sous jacent, ou par la méthode de comparaison des données VMS avec les données scientifiques. Dans le cas présenté ici, une étude faite en utilisant la densité d'utilisation conditionnelle estimée par la méthode des Ponts Browniens a montré des résultats similaires.
Nous pensons que la question de l'information sur la ressource ciblée que contient la donnée VMS est encore ouverte et mérite une étude plus approfondie, car elle pourrait être d'un enjeu majeur pour l'évaluation de la ressource.

4.2.2 Perspectives en écologie du mouvement pour l'utilisation des EDS dérivant d'un potentiel

Un potentiel induisant un processus stationnaire

Comme déjà développé dans la section 3.4, le modèle GaP ne décrit pas un processus ergodique, et ainsi ne conduit pas à une distribution stationnaire pour le champ spatial de la probabilité de présence des individus. Ainsi, la distribution spatiale d'un individu ne se stabilise pas au cours du temps. D'un point de vue écologique et halieutique, proposer des modèles stationnaires (ou qui tendent vers une distribution stationnaire) rendrait leur interprétation plus intuitive.
Dans la formulation du modèle proposée, l'absence de distribution stationnaire provient du caractère borné de la fonction de potentiel proposée. Plusieurs options sont possibles pour relaxer cette hypothèse.
Une première possibilité serait de décrire le mouvement comme un processus de Ornstein Ulhenbeck. Malheureusement, ce modèle découle d'une carte de potentiel unimodal. Or, nous l'avons vu sur les données étudiées, l'idée d'une seule zone attractive pour résumer le mouvement est sans doute trop restrictive.
Une extension du modèle GaP, qui garderait la même structure, consisterait à développer la fonction de potentiel P afin de la rendre non bornée. La nouvelle fonction de potentiel $\tilde{P}$ dont découlerait le mouvement serait alors

$$\tilde{P}(x) = P(x) - c\sqrt{(1+ \parallel x - \nu \parallel^2)}$$

où P est de la forme GaP (équation (3.24)), le terme c est une constante réelle (intuitivement, une force de rappel, qu'il faudrait estimer) et ν est un centre attractif (qui pourrait être le barycentre des zones $(\mu_k)_{k=1,\dots,K}$ par exemple, ou supposé connu). Cette forme aurait un double avantage :

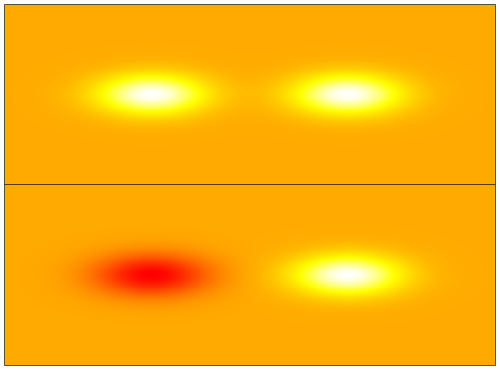


FIGURE *4.2 – Exemple de fonctions de potentiel régissant le modèle GaP avec 2 modes. En haut, le modèle GaP actuel avec deux zones attractives (en blanc). En bas, le signe du poids π_1 pour le mode de gauche a été changé. La zone devient alors une zone répulsive. Ce changement dans le modèle est négligeable pour l'inférence*

– Elle préserve l'existence d'une unique solution de l'EDS

$$dX_t = \nabla \tilde{P}(X)dt + \gamma dW_t$$

et, comme dit plus haut, cette solution aurait alors un distribution stationnaire.

– Ce processus garde les propriétés du modèle GaP concernant l'estimation. Ainsi toute la méthode d'estimation proposée ici, se basant sur l'algorithme de simulation conditionnelle exacte EA1, reste valable. L'estimation par algorithme MCEM des paramètres du modèle suit alors la même procédure que celle décrite dans la section 3.2. La seule différence réside dans le calcul des bornes (D.6), (D.7), et (D.8), nécessaires à la simulation conditionnelle.

Nous proposons cette piste d'amélioration comme une piste privilégiée pour de futurs développements du modèle GaP.

Un potentiel comprenant des zones répulsives

Comme amorcé dans la section 3.4, une autre modification intéressante du modèle GaP consiste à introduire des zones répulsives dans le potentiel P. En effet, il est tout à fait envisageable que certaines zones de l'environnement aient un caractère répulsif pour un individu. Par exemple, des zones où les températures sont trop éloignées de la niche environnementale des individus peuvent être répulsives. En halieutique, des zones à fonds rocheux sont répulsives pour des chalutiers de fond.
Une proposition naturelle est d'écrire le même modèle que celui proposé en équation (3.24), mais en relaxant la contrainte de positivité des poids. Une telle carte de potentiel est représentée sur la figure 4.2. Cette modification n'impacterait aucunement les conditions d'existence d'une solution à l'EDS correspondante. Du point de vue de l'inférence, ce changement est immédiat. Il demande simplement un nouveau calcul des bornes (D.6), (D.7), et (D.8). L'estimation par algorithme MCEM des paramètres du modèle suit alors la même procédure que celle décrite dans la section 3.2.

Un potentiel comprenant des zones interdites ou inaccessibles

Pour un individu il est des zones qui sont inaccessibles (une zone émergée pour un poisson) ou interdites (une zone réglementaire pour un navire de pêche). Pour représenter une telle zone dans un champ de potentiel, une zone interdite serait alors une zone au potentiel arbitrairement faible (voire infiniment faible). Pour conserver le formalisme du modèle GaP, il faudrait cependant que la fonction P comprenant de telles zones soit dérivable deux fois. Mais cela revient alors à modéliser les zones interdites comme des zones répulsives du paragraphe précédent. Cependant une telle modélisation ne rend pas compte de l'existence de frontières nettes entre les zones où la présence est possible, et celles où la présence est impossible. L'incorporation d'une frontière dans le potentiel pose cependant un problème majeur. En effet, si la dérive de l'EDS $b(x) := \nabla P(x)$ n'est pas continue alors, la condition de Lipschitz

$$\exists C \text{ tel que } \forall x, y \in \mathbb{R}^2, \parallel b(x) - b(y) \parallel \leq C \parallel x - y \parallel$$

n'est plus satisfaite. Ainsi, la condition d'existence d'une solution de notre EDS n'est plus satisfaite. Une éventuelle solution serait alors de réduire le domain de l'EDS aux zones possibles, et d'introduire des frontières réflectives pour les zones interdites. L'atteinte d'une frontière pousse à "rebondir" dans la direction opposée à cette frontière. Pour le mouvement Brownien, une telle possibilité existe, donnant lieu au processus du mouvement Brownien réfléchi (Dieker, 2010). Cependant, nous ignorons si une extension au modèle GaP est possible. Si tel était le cas, nous ignorons également dans quelle mesure la méthode d'inférence proposée dans ce manuscrit resterait valable pour un mouvement réfléchi. De tels développements offrent donc des perspectives de recherche intéressantes.

Un potentiel affecté par l'environnement

Un des avantages de l'approche développée dans le chapitre 3 est le caractère linéaire de l'opérateur gradient. Cette caractéristique ouvre la voie à une décomposition du potentiel dont dérive le mouvement sous la forme d'une combinaison de différents potentiels aux origines diverses.

Ainsi, supposons que l'on dispose de deux cartes connues, une carte P_A définissant l'abondance en proies et une carte P_E décrivant une certaine caractéristique environnementale. Supposons qu'on ait accès au gradient de ces cartes en tout point de l'espace. Supposons également que l'individu ait une perception de l'espace personnelle, inconnue, due à d'autres facteurs, notée P_P. Cette dernière carte peut être vue comme une sorte d'analogue d'un effet individuel, par rapport aux cartes P_A et P_E, communes aux individus. Alors, un modèle étendu décrivant le déplacement d'un individu en fonction de ces trois cartes pourrait être de la forme :

$$\begin{aligned} dX_t &= \nabla \left(P_P(X_t) + \alpha P_A(X_t) + \beta P_E(X_t) \right) + \gamma dW_t \\ &= \nabla P_P(X_t) + \alpha \nabla P_A(X_t) + \beta \nabla P_E(X_t) + \gamma dW_t \end{aligned} \tag{4.1}$$

Où α et β serait les poids relatifs de chaque carte (positifs ou négatifs) qu'il conviendrait d'estimer, conjointement avec la carte P_P et le paramètre γ. Encore une fois, la méthode générique d'estimation proposée dans cette thèse pourrait être facilement transposée à un tel modèle, au prix de quelques changements mineurs. En effet, la fonction définie dans (3.29) reste peu changée et calculable en tous points. Les bornes nécessaires à l'utilisation de l'algorithme exact restent également calculables.
Ce modèle pourrait permettre de tester l'influence de covariables dans le déplacement d'un individu, et de décomposer les facteurs influençant les trajectoires sous la formes de facteurs communs à plusieurs individus, et d'effets individuels. L'utilisation de la méthode d'inférence proposée dans la section 3.2, demande cependant une connaissance suffisamment fine de la distribution de la covariable pour avoir accès à un gradient (ou une estimation de celui ci) continu. L'idéal est donc de disposer d'une expression fonctionnelle de la distribution de la covariable.

Un potentiel dérivant des interactions entre individus

L'émergence des domaines vitaux dérivent, pour beaucoup d'animaux, des interactions avec leur congénères (Giuggioli *et al.*, 2011; Giuggioli and Kenkre, 2014). Ainsi, le mouvement de certains animaux est grandement influencé par ces interactions. Ces interactions peuvent être coopératives ou de compétition. Réussir à quantifier ces comportements d'interactions est un problème ouvert aujourd'hui en écologie (Long *et al.*, 2014). En halieutique, pour étudier la dynamique des flottilles, évaluer les interactions entre les pêcheurs est un défi important pour comprendre la structuration de l'activité de pêche. De même évaluer l'interaction entre les navires de pêches et le trafic maritime est un enjeu important pour la bonne définition de mesures de gestion. Le formalisme présenté ici a déjà été proposé par Brillinger *et al.* (2011) pour représenter les interactions entre individus.

Dans le cas de deux individus se déplaçant simultanément dans un même espace, on peut s'intéresser à représenter conjointement la trajectoire de ces deux individus. Ces deux trajectoires sont donc représentées dans un système d'EDS, guidées par une fonction de potentiel faisant intervenir les interactions entre les positions des deux individus :

$$dX_t^{(1)} = \nabla P_1(X_t^{(1)}, X_t^{(2)}) + \gamma dW_t, \quad X_0^{(1)} = X_0^{(1)} \tag{4.2}$$
$$dX_t^{(2)} = \nabla P_2(X_t^{(1)}, X_t^{(2)}) + \gamma dW_t, \quad X_0^{(2)} = X_0^{(2)} \tag{4.3}$$

Dans ce système, le déplacement d'un individu ne dépend pas seulement de sa position dans l'espace, mais aussi de celle de l'autre individu. Dans un premier temps, il serait tout à fait envisageable de séparer cette fonction de potentiel conjointe sous la forme d'une somme entre un potentiel commun P, et d'une interaction V, tels que le système décrit par (4.2) et (4.3)

deviendrait

$$dX_t^{(1)} = \nabla \left(P(X_t^{(1)}) + V(X_t^{(1)}, X_t^{(2)}) \right) + \gamma dW_t, \quad X_0^{(1)} = X_0^{(1)} \tag{4.4}$$

$$dX_t^{(2)} = \nabla \left(P(X_t^{(2)}) + V(X_t^{(1)}, X_t^{(2)}) \right) + \gamma dW_t, \quad X_0^{(2)} = X_0^{(2)} \tag{4.5}$$

Ainsi, ce modèle nécessiterait une paramétrisation adéquate d'une fonction d'interaction. L'estimation de paramètres permettrait alors de déterminer si deux individus ont tendance à s'attirer, ou à s'éviter. Le modèle ci dessus pourrait être étendu à n individus.
Dans leur article, Brillinger *et al.* (2011), pour estimer les paramètres, utilisaient un schéma d'Euler, dont la fiabilité est grandement dépendante de la haute fréquence de l'échantillonnage. Nous pensons qu'il est possible, pour le modèle défini par (4.4) et (4.5), d'utiliser les algorithmes d'estimation exacte utilisés dans la section 3.2.

Un potentiel inhomogène en temps

Comme dit dans la section 4.1 dans le cadre des HMM, l'hypothèse d'homogénéité du potentiel semble peu compatible avec la description du mouvement d'individus ayant des trajectoires marquées par un retour à leur point de départ. C'est notamment le cas des navires de pêche qui, très souvent, rejoignent leur port d'attache après chaque marée. Ainsi, un navire part d'un port (qui est alors un point répulsif) pour y revenir (qui est devenu un point attractif). Il faudrait donc que le modèle de mouvement réussisse à intégrer des changements temporels dans la forme du potentiel guidant le mouvement.
Ces changements pourraient advenir de manière continue, au travers d'une carte de potentiel $P(x, t)$. Cette formulation demanderait une explicitation de la dépendance au temps. Elle pourrait traduire l'épuisement de la ressource à l'intérieur d'une zone donnée, durant la trajectoire.
Une alternative, pour laquelle nous avons une préférence, serait l'intégration d'un modèle de mouvement de type GaP, à un modèle de comportement de type HMM, comme développé ci-dessous

4.3 Modéliser et reconstruire des comportements et des potentiels ?

Comme nous l'avons dit plus haut, le modèles HMM ont pour objectif principal de reconstruire la séquence des comportements au cours de la trajectoire. Ainsi, dans ces modèles, la description du mouvement y est souvent faite en termes de marches aléatoires corrélées en temps discrets, conditionnellement au comportement, lui même modélisé au travers d'une chaîne de Markov.
On modélise, dans un schéma discret, la longueur et la direction des pas successifs (cette modélisation est conjointe dans le modèle AR HMM). Or, tel que défini, aucune cause du déplacement n'est concrètement exprimée dans le modèle. En particulier, la marche aléatoire d'un individu

conditionnellement à son comportement est indépendante de sa position dans l'espace.
D'un autre côté, le modèle de mouvement continu développé au chapitre 3 ne décrit que le mouvement, dans un cadre spatialisé, mais selon des mécanismes homogènes dans le temps, ne permettant pas de décrire la succession des comportements au cours d'une trajectoire.

Ainsi, une perspective naturelle à ce travail de thèse consister à coupler deux processus, un processus comportemental, et un mouvement spatialisé. Les deux processus pourraient être ceux décrit dans le corps des chapitres 2 et 3, ou des extensions proposées dans les sections 4.1 et 4.2.
Une difficulté majeure de tels développements est la différence de formalisation du temps entre les deux approches. Dans ce manuscrit, l'approche comportementale s'est faite par une formulation discrète du temps. Cette formulation s'expliquait par la nature du problème posé [3]. Or il est possible de formuler une chaîne de Markov en temps continu (Norris, 1998b), et d'intégrer ce formalisme à un modèle de Markov Caché. Cette approche a récemment été adoptée pour un modèle décrivant les activités de navires de pêche (Charles *et al.*, 2014).
Récemment, cette approche de couplage entre un modèle comportemental et une EDS a donné lieu à une modèle de mouvement en écologie animale (Harris and Blackwell, 2013). Ce modèle en temps continu suppose qu'un individu adopte plusieurs comportements. Conditionnellement à un comportement, la trajectoire de l'individu est supposée être la réalisation d'un processus de Ornstein Ulhenbeck, dont les paramètres dépendent du comportement :

$$dX_t|\{S_t = i\} = \rho_i(\mu_i - X_t)dt + \sigma_i dW_t\,.$$

Les transitions entre comportements sont alors supposées dépendre de l'habitat, et non nécessairement homogènes dans le temps. La méthode d'estimation des paramètres de ce modèles a été publiée très [4] récemment (Blackwell *et al.*, 2015). Comme dans la méthode présentée en section 3.2, l'estimation se base sur la capacité à simuler exactement le processus solution de l'EDS avec saut. Cette possibilité est facilitée par les propriétés remarquables du processus de Ornstein Ulhenbeck, qui est un processus Gaussien. Cette simulation permet l'estimation des distributions a posteriori des paramètres et des comportements, dans un cadre Bayésien, utilisant des algorithmes de type Monte Carlo Markov Chains (MCMC).
Le cadre proposé par le modèle de Harris and Blackwell (2013) constitue un pas très intéressant vers des modèles plus riches capables de coupler, dans un formalisme en temps continu, un processus caché de changement d'état comportemental avec un modèle de mouvement stochastique, dont les paramètres restent estimables à partir de données de trajectoire.

3. Mais aussi par la peur de l'auteur pour le temps continu à l'époque de l'étude !
4. Trop !

4.4 Conclusion générale

4.4.1 Vers des modèles prédictifs ?

Il nous semble que la description du mouvement comme la résultante d'un potentiel constitue une base fondamentale adaptée pour bâtir des modèles prédictifs de mouvement. En effet, des scénarios spatialisés peuvent se retranscrire sous la forme d'une redéfinition des potentiels. En supposant que les mécanismes du déplacement restent similaires, on pourrait alors évaluer, sous ce nouveau potentiel, la redistribution moyenne des individus dans l'espace, et les conséquences éventuelles de cette redistribution.
L'obtention de modèles prédictifs pour le déplacement des individus est un défi enthousiasmant dans une optique d'aide à la gestion, et ce dans de multiples applications :

- En halieutique, un objectif essentiel est de disposer d'outils permettant d'éclairer les décisions de gestion (fermeture d'une zone, gestion d'un stock) aux travers d'une modélisation de la réponse des pêcheurs et des flottilles. Des modèles prédictifs adaptés de mouvement des navires de pêches adaptés pourraient ainsi servir d'outil dans ce cadre.
- En écologie animale, des modèles prédictifs adaptés de mouvement pourraient permettre de :
    1. Mieux définir l'aire de répartition de certaines espèces ;
    2. Mieux comprendre la distribution de certaines espèces dans un nouvel environnement (espèces invasives) ;
    3. Anticiper les impacts de transformation anthropiques de l'environnement sur la distribution d'espèces.

4.4.2 Le mot de la fin

Si les modèles présentés dans ce manuscrit sont encore loin de modèles prédictifs, leur élaboration a permis de se questionner sur la formulation adéquate pour les mécanismes du mouvement. Nous avons pu par exemple nous interroger sur la formulation continue du temps, dont la pertinence est discutée en écologie du mouvement (McClintock *et al.*, 2014). Au vu de ces travaux, cette formulation nous apparait adaptée pour décrire le mouvement d'un individu, car elle permet, entre autres, de s'affranchir d'hypothèses sur l'échantillonnage, et d'être en adéquation avec la nature continue de la trajectoire.
Les modèles décrits utilisent des briques essentielles à la construction de modèles prédictifs adéquats décrivant le mouvement. Ainsi, les perspectives proposées se basent sur ces outils, et ne demandent pas nécessairement d'outils d'inférence nouveaux pour atteindre des objectifs d'intégration de nouveau mécanismes, comme les interactions, ou la non homogénéité des transitions comportementales.

Les travaux menés dans cette thèse ont aussi montré tout l'intérêt de décloisonner des domaines d'application différents comme celui de la dynamique des flottilles de pêche en halieutique, et le

vaste domaine de l'écologie du mouvement en écologie animale. Cette dernière application est à l'origine de travaux récents, novateurs et enthousiasmants (très subjectivement : Giuggioli and Kenkre (2014); Fleming *et al.* (2015); Blackwell *et al.* (2015)), dont les problématiques sont souvent très similaires à celles de l'halieutique.

Les données VMS utilisées dans ce travail représentent une base de données abondante. À ce jour, l'accessibilité à ces données demeure difficile (pour des raisons réglementaires et de confidentialité), ainsi, relaxer ces contraintes d'accessibilité permettrait à une large communauté scientifique de bénéficier de cette source de données unique. En effet, cette base est unique en écologie du mouvement au sens où elle couvre l'ensemble d'une population sur une zone géographique donnée, sur un temps long. Ces données ont ainsi un formidable potentiel pour le développement de modèles de mouvements, tant pour l'halieutique, que pour l'écologie du mouvement en général.

Les briques construites dans ce manuscrit sont malheureusement encore trop molles pour soutenir une telle étude. Cependant, nous espérons que ces travaux, qui ont ouvert plus de questions qu'ils n'ont apporté de réponses, pourront servir de base à de futures recherches pour la modélisation de trajectoires en halieutique et en écologie du mouvement.

Bibliographie

Aït-Shalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics*, 36(2) :906–937.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, pages 610–624. Springer.

Andre, C., Franck, C., Lucie, C., Jean-Claude, D., Juliette, D., Jean-Marie, D., Ludovic, D., Aurélie, F., Clément, G., Laure, G., Stuart, H., Roger, J., Philippe, K., Valentina, L., Corinne, M., Geoff, M., Jocelyne, M., Yoshi, O., Emilie, R., Bob, S., Nicolas, S., Sandrine, V., Ching-Maria, V., Yves, V., Joanne, W., and Caroline, W. (2009). Atlas des Habitats des Ressources Marines de la Manche Orientale - CHARM II.

Aristote. Le mouvement des animaux ; suivi de la locomotion des animaux.

Avgar, T., Mosser, A., Brown, G. S., and Fryxell, J. M. (2013). Environmental and individual drivers of animal movement patterns across a wide geographical gradient. *Journal of Animal Ecology*, 82(1) :96–106.

Barraquand, F. and Benhamou, S. (2008). Animal movements in heterogeneous landscapes : identifying profitable places and homogeneous bouts. *Ecology*, 89(12) :3336–3348.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1) :164–171.

Benhamou, S. (2011). Dynamic approach to space and habitat use based on biased random bridges. *PloS one*, 6 :1–8.

Benoit-Bird, K. J. and Au, W. W. (2003). Prey dynamics affect foraging by a pelagic predator (stenella longirostris) over a range of spatial and temporal scales. *Behavioral Ecology and Sociobiology*, 53(6) :364–373.

Berthou, P., Bégot, E., Laurans, M., Campéas, A., Leblond, E., and Habasque, J. (2013). Présentation de la suite logicielle AlgoPesca. Rapport interne Ifremer.

Bertrand, S., Bertrand, A., Guevara-Carrasco, R., and Gerlotto, F. (2007). Scale-invariant movements of fishermen : The same foraging strategy as natural predators. *Ecological Applications*, 17(2) :331–337.

Bertrand, S., Diaz, E., and Lengaigne, M. (2008). Patterns in the spatial distribution of peruvian anchovy (engraulis ringens) revealed by spatially explicit fishing data. *Progress in Oceanography*, 79(2-4) :379–389.

Bertrand, S., Diaz, E., and Niquen, M. (2004). Interactions between fish and fisher's spatial distribution and behaviour : an empirical study of the anchovy (*Engraulis ringens*) fishery of peru. *ICES Journal of Marine Science*, 61(7) :1127–1136.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. (2006a). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6) :1077–1098.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. (2009). Monte Carlo maximum likelihood estimation for discretely observed diffusions processes. *Annals of Statistics*, 37(1) :223–245.

Beskos, A., Papaspiliopoulos, O., Roberts, G., and Fearnhead, P. (2006b). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of Royal Statistical Society, Series B : Statistical Methodology*, 68 :333–382.

Beskos, A. and Roberts G., O. (2005). Exact simulation of diffusions. *Annals of Applied Probability*, 15(4) :2422–2444.

Bez, N., Mahévas, S., Etienne, M.-P., Monestiez, P., Rivot, E., Gloaguen, P., Vermard, Y., Woillez, M., Bertrand, S., Joo, R., Delattre, M., Nerini, D., Walker, E., and De Pontual, H. (In prep). Revisiting markov state space models with validation data.

Bez, N., Walker, E., Gaertner, D., Rivoirard, J., and Gaspar, P. (2011). Fishing activity of tuna purse seiners estimated from vessel monitoring system (VMS) data. *Canadian Journal of Fisheries and Aquatic Sciences*, 68(11) :1998–2010.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7) :719–725.

Biseau, A. (1998). Definition of a directed fishing effort in a mixed-species trawl fishery, and its impact on stock assessments. *Aquatic Living Resources*, 11(3) :119–136.

Blackwell, P. (1997). Random diffusion models for animal movement. *Ecological Modelling*, 100(1–3) :87 – 102.

Blackwell, P. G., Niu, M., Lambert, M. S., and LaPoint, S. D. (2015). Exact bayesian inference for animal movement in continuous time. *Methods in Ecology and Evolution*.

Bovet, P. and Benhamou, S. (1988). Spatial-analysis of animals movements using a correlated random-walk model. *Journal of Theoretical Biology*, 131(4) :419–433.

Boyce, M. S. and McDonald, L. L. (1999). Relating populations to habitats using resource selection functions. *Trends in Ecology & Evolution*, 14(7) :268–272.

Breed, G. A., Costa, D. P., Jonsen, I. D., Robinson, P. W., and Mills-Flemming, J. (2012). State-space methods for more completely capturing behavioral dynamics from animal tracks. *Ecological Modelling*, 235 :49–58.

Brillinger, D. (2010). *Handbook of Spatial Statistics*, chapter 26. Chapman and Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Brillinger, D., Preisler, H., Wisdom, M., *et al.* (2011). Modelling particles moving in a potential field with pairwise interactions and an application. *Brazilian Journal of Probability and Statistics*, 25(3) :421–436.

Brillinger, D. R., Preisler, H. K., Ager, A. A., and Kie, J. (2001a). The use of potential functions in modelling animal movement. *Data analysis from statistical foundations*, pages 369–386.

Brillinger, D. R., Preisler, H. K., Ager, A. A., Kie, J., and Stewart, B. S. (2001b). Modelling movements of free-ranging animals. *Univ. Calif. Berkeley Statistics Technical Report*, 610.

Brown, R. (1828). A brief account of microscopical observations made in the months of june, july and august, 1827, on the particles contained in the pollen of plants ; and on the general existence of active molecules in organic and inorganic bodies.

Bullard, F. (1999). *Estimating the home range of an animal, a Brownian Bridge approach.* PhD thesis, University of North Carolina at Chapell Hill.

Burt, W. H. (1943). Territoriality and home range concepts as applied to mammals. *Journal of Mammalogy*, 24(3) :pp. 346–352.

Cagnacci, F., Boitani, L., Powell, R. A., and Boyce, M. S. (2010). Animal ecology meets gps-based radiotelemetry : a perfect storm of opportunities and challenges. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1550) :2157–2162.

Calenge, C. (2006). The package "adehabitat" for the r software : A tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197(3–4) :516 – 519.

Campbell, M. S., Stehfest, K. M., Votier, S. C., and Hall-Spencer, J. M. (2014). Mapping fisheries for marine spatial planning : Gear-specific vessel monitoring system (VMS), marine conservation and offshore renewable energy. *Marine Policy*, 45 :293–300.

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2014). A robust approach to multiple change-point estimation in an AR(1) process. *arXiv preprint arXiv :1403.1958*.

Chang, S.-K. (2011). Application of a vessel monitoring system to advance sustainable fisheries management—Benefits received in Taiwan. *Marine Policy*, 35(2) :116–121.

Charles, C., Gillis, D., and Wade, E. (2014). Using hidden markov models to infer vessel activities in the snow crab (chionoecetes opilio) fixed gear fishery and their application to

catch standardization. *Canadian Journal of Fisheries and Aquatic Sciences*, 71(12) :1817–1829.

Chavez, A. S. G. E. M. (2006). Landscape Use and Movements of Wolves in Relation to Livestock in a Wildland–Agriculture Matrix. *Journal of Wildlife Management*, 70(4) :1079–1086.

Codling, E. and Hill, N. (2005). Sampling rate effects on measurements of correlated and biased random walks. *Journal of Theoretical Biology*, 233(4) :573–588.

Codling, E. A., Plank, M. J., and Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal Society Interface*, 5(25) :813–834.

Cooke, S., Hinch, S., Wikelski, M., Andrews, R., Kuchel, L., Wolcott, T., and Butler, P. (2004). Biotelemetry : a mechanistic approach to ecology. *Trends in Ecology and Evolution*, 19(6) :334–343.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics*, pages 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1) :1–38 (with discussion).

Dieker, A. (2010). Reflected brownian motion. *Wiley Encyclopedia of Operations Research and Management Science*.

Ding, L., Fan, H., and Meng, L. (2015). Understanding taxi driving behaviors from movement data. In Bacao, F., Santos, M. Y., and Painho, M., editors, *Agile 2015*, Lecture Notes in Geoinformation and Cartography, pages 219–234. Springer International Publishing.

Ditlevsen, S. and Samson, A. (2014). Estimation in the partially observed stochastic morris–lecar neuronal model with particle filter and stochastic approximation methods. *The Annals of Applied Statistics*, 8(2) :674–702.

Dubé, B., Diméet, J., Rochet, M. J., Tétard, A., Gaudou, O., Messanot, C., Biseau, A., , and Salaün, M. (2012). Observations à bord des navires de pêche professionnelle. bilan de l'échantillonnage 2011.

Dunn, J. E. and Gipson, P. S. (1977). Analysis of radio telemetry data in studies of home range. *Biometrics*, 33(1) :pp. 85–101.

Dunn, M. (1999). *The exploitation of selected non-quota species in the English Channel*. PhD thesis, University of Portsmouth.

Durham, G. and Gallant, A. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes (with discussion). *J. Bus. Econom. Statist.*, 20 :297–338.

Eddelbuettel, D. and François, R. (2011). Rcpp : Seamless R and C++ integration. *Journal of Statistical Software*, 40(8) :1–18.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics & Applied Probability. Chapman & Hall.

Einstein, A. (1906). On the theory of the brownian movement. *Annalen der physik*, 4(19) :371–381.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69 :959–993.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *J. Bus. Econom. Statist.*, 19 :177–191.

Evans, M. R., Norris, K. J., and Benton, T. G. (2012). Predictive ecology : systems approaches. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 367(1586) :163–169.

Firle, S., Bommarco, R., Ekbom, B., and Natiello, M. (1998). The influence of movement and resting behavior on the range of three carabid beetles. *Ecology*, 79(6) :2113–2122.

Fleming, C. H., Fagan, W. F., Mueller, T., Olson, K. A., Leimgruber, P., and Calabrese, J. M. (2015). Rigorous home range estimation with movement data : a new autocorrelated kernel density estimator. *Ecology*, 96(5) :1182–1188.

Flemming, J. E. M., Field, C. A., James, M. C., Jonsen, I. D., and Myers, R. A. (2006). How well can animals navigate ? estimating the circle of confusion from tracking data. *Environmetrics*, 17(4) :351–362.

Fort, G. and Moulines, E. (2003). Convergence of the monte carlo expectation maximization for curved exponential families. *Annals of Statistics*, 31(4) :1220–1259.

Franke, A., Caelli, T., Kuzyk, G., and Hudson, R. J. (2006). Prediction of wolf (canis lupus) kill-sites using hidden markov models. *Ecological Modelling*, 197(1) :237–246.

Freeman, R., Dean, B., Kirk, H., Leonard, K., Phillips, R. A., Perrins, C. M., and Guilford, T. (2013). Predictive ethoinformatics reveals the complex migratory behaviour of a pelagic seabird, the manx shearwater. *Journal of the Royal Society Interface*, 10(84).

Fretwell, S. D. (1972). *Populations in a seasonal environment*. Number 5. Princeton University Press.

Fryxell, J. M., Hazell, M., Borger, L., Dalziel, B. D., Haydon, D. T., Morales, J. M., McIntosh, T., and Rosatte, R. C. (2008). Multiple movement modes by large herbivores at multiple spatiotemporal scales. *Proceedings of the National Academy of the United States of America*, 105(49) :19114–19119.

Gerritsen, H. and Lordan, C. (2011). Integrating vessel monitoring systems (VMS) data with daily catch data from logbooks to explore the spatial distribution of catch and effort at high resolution. *ICES Journal of Marine Science*, 68(1) :245–252.

Getz, W. M. and Saltz, D. (2008). A framework for generating and analyzing movement paths on ecological landscapes. *Proceedings of the National Academy of the United States of America*, 105(49) :19066–19071.

Gillis, D. M. (2003). Ideal free distributions in fleet dynamics : a behavioral perspective on vessel movement in fisheries analysis. *Canadian Journal of Zoology*, 81(2) :177–187.

Girardin, R. (2015). *Ecosystem and fishers' behabiour modelling : two crucial and interacting approaches to support Ecosystem Based Fisheries Management in the Eastern Channel*. PhD thesis, Université Lille 1.

Girardin, R., Vermard, Y., Thebaud, O., Tidd, A., and Marchal, P. (2015). Predicting fisher response to competition for space and resources in a mixed demersal fishery. *Ocean & Coastal Management*, 106 :124–135.

Giuggioli, L. and Kenkre, V. (2014). Consequences of animal interactions on their dynamics : emergence of home ranges and territoriality. *Mov Ecol*, 2 :20.

Giuggioli, L., Potts, J. R., and Harris, S. (2011). Animal interactions and the emergence of territoriality.

Gloaguen, P., Mahévas, S., Rivot, E., Woillez, M., Guitton, J., Vermard, Y., and Etienne, M. P. (2015). An autoregressive model to describe fishing vessel movement and activity. *Environmetrics*, 26(1) :17–28.

Guédon, Y. (2003). Estimating hidden semi-markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3) :604–639.

Guilford, T., Roberts, S., Biro, D., and Rezek, L. (2004). Positional entropy during pigeon homing ii : navigational interpretation of bayesian latent state models. *Journal of Theoretical Biology*, 227(1) :25–38.

Guinet, C., Vacquié-Garcia, J., Picard, B., Bessigneul, G., Lebras, Y., Dragon, A.-C., Viviant, M., Arnould, J. P., Bailleul, F., *et al.* (2014). Southern elephant seal foraging success in relation to temperature and light conditions : insight into prey distribution. *Mar. Ecol. Prog. Ser*, 499 :285–301.

Gurarie, E. (2008). *Models and analysis of animal movements : From individual tracks to mass dispersal*. PhD thesis, University of Washington.

Gurarie, E., Andrews, R. D., and Laidre, K. L. (2009). A novel method for identifying behavioural changes in animal movement data. *Ecology Letters*, 12(5) :395–408.

Hansen, N. and Kern, S. (2004). Evaluating the CMA Evolution Strategy on Multimodal Test Functions. *Eighth International Conference on Parallel Problem Solving from Nature*, 72 :337–354.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2) :159–195.

Harris, K. J. and Blackwell, P. G. (2013). Flexible continuous-time modelling for heterogeneous animal movement. *Ecological Modelling*, 255 :29–37.

Hayne, D. W. (1949). Calculation of size of home range. *Journal of Mammalogy*, 30(1) :pp. 1–18.

Hinch, S., Standen, E., Healey, M., and Farrell, A. (2002). Swimming patterns and behaviour of upriver-migrating adult pink (oncorhynchus gorbuscha) and sockeye (o. nerka) salmon as assessed by emg telemetry in the fraser river, british columbia, canada. *HYDROBIOLOGIA*, 483(1-3) :147–160.

Hintzen, N. T., Bastardie, F., Beare, D., Piet, G. J., Ulrich, C., Deporte, N., Egekvist, J., and Degel, H. (2012). VMStools : Open-source software for the processing, analysis and visualisation of fisheries logbook and VMS data. *Fisheries Research*, 115 :31–43.

Hintzen, N. T., Piet, G. J., and Brunel, T. (2010). Improved estimation of trawling tracks using cubic Hermite spline interpolation of position registration data . *Fisheries Research*, 101(1–2) :108 – 115.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, 13(3) :325–347.

Horne, J. S., Garton, E. O., Krone, S. M., and Lewis, J. S. (2007). Analyzing animal movements using brownian bridges. *Ecology*, 88(9) :2354–2363.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 48(1) :15–30.

Hutton, T., Mardle, S., Pascoe, S., and Clark, R. (2004). Modelling fishing location choice within mixed fisheries : English north sea beam trawlers in 2000 and 2001. *ICES Journal of Marine Science*, 61(8) :1443–1452.

Iacus, S. M. (2009). *Simulation and inference for stochastic differential equations : with R examples*, volume 1. Springer Science & Business Media.

Johnson, D. S., London, J. M., Lea, M.-A., and Durban, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology*, 89(5) :1208–1215.

Jonsen, I., Myers, R., and Flemming, J. (2003). Meta-analysis of animal movement using state-space models. *Ecology*, 84(11) :3055–3063.

Jonsen, I. D., Basson, M., Bestley, S., Bravington, M. V., Patterson, T. A., Pedersen, M. W., Thomson, R., Thygesen, U. H., and Wotherspoon, S. J. (2013a). State-space models for bio-loggers : A methodological road map. *Deep Sea Research Part II- Topical Studies in Oceanography*, 88-89(SI) :34–46.

Jonsen, I. D., Basson, M., Bestley, S., Bravington, M. V., Patterson, T. A., Pedersen, M. W., Thomson, R., Thygesen, U. H., and Wotherspoon, S. J. (2013b). State-space models for bio-loggers : A methodological road map. *Deep Sea Research Part II Topical Studies in Oceanography*, 88-89(SI) :34–46.

Jonsen, I. D., Flenming, J. M., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology*, 86(11) :2874–2880.

Jonsen, I. D., Myers, R. A., and James, M. C. (2006). Robust hierarchical state-space models reveal diel variation in travel rates of migrating leatherback turtles. *Journal of Animal Ecology*, 75(5) :1046–1057.

Jonsen, I. D., Myers, R. A., and James, M. C. (2007). Identifying leatherback turtle foraging behaviour from satellite telemetry using a switching state-space model. *Marine Ecology-Progress Series*, 337 :255–264.

Joo, R. (2013). *A behavioral ecology of fishermen : hidden stories from trajectory data in the Northern Humboldt Curren System.* PhD thesis, Université de Montpellier II.

Joo, R., Bertrand, S., Chaigneau, A., and Niquen, M. (2011). Optimization of an artificial neural network for identifying fishing set positions from VMS data : An example from the peruvian anchovy purse seine fishery. *Ecological Modelling*, 222(4) :1048–1059.

Joo, R., Bertrand, S., Tam, J., and Fablet, R. (2013a). Hidden markov models : The best models for forager movements ? *Plos One*, 8(8).

Joo, R., Sophie, B., Jorge, T., and Ronan, F. (2013b). Hidden Markov Models : The Best Models for Forager Movements ? *PLoS ONE*, 8(8) :e71246.

Keating, K. A. and Cherry, S. (2009). Modeling utilization distributions in space and time. *Ecology*, 90(7) :1971–1980.

Kie, J. G., Matthiopoulos, J., Fieberg, J., Powell, R. A., Cagnacci, F., Mitchell, M. S., Gaillard, J.-M., and Moorcroft, P. R. (2010). The home-range concept : are traditional estimators still relevant with modern telemetry technology ? *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1550) :2221–2231.

Kiester, A. and Slatkin, M. (1974). A strategy of movement and resource utilization. *Theoretical population biology*, 6(1) :1–20.

Kourti, N., Shepherd, I., Greidanus, H., Alvarez, M., Aresu, E., Bauna, T., Chesworth, J., Lemoine, G., and Schwartz, G. (2005). Integrating remote sensing in fisheries control. *Fisheries Management and Ecology*, 12(5) :295–307.

Kranstauber, B., Kays, R., LaPoint, S. D., Wikelski, M., and Safi, K. (2012). A dynamic brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement. *Journal of Animal Ecology*, 81(4) :738–746.

Kranstauber, B., Safi, K., and Bartumeus, F. (2014). Bivariate gaussian bridges : directional factorization of diffusion in brownian bridge models. *Movement Ecology*, 2(1).

Krige, D. (1951). *A statistical approach to some mine valuation and allied problems on the Witwatersrand : By DG Krige*. PhD thesis, University of the Witwatersrand.

Kwan, M. (1998). Space-time and integral measures of individual accessibility : A comparative analysis using a point-based framework. *Geographical Analysis*, 30(3) :191–216.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data : hidden markov models and extensions. *Ecology*, 93(11) :2336–2342.

Langrock, R. and Zucchini, W. (2011). Hidden markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1) :715–724.

Laukkanen, M. (2003). Cooperative and non-cooperative harvesting in a stochastic sequential fishery. *Journal of Environmental Economics and Management*, 45(2, Supplement) :454 – 473.

Laver, P. N. and Kelly, M. J. (2008). A critical review of home range studies. *Journal of the Wildlife Management*, 72(1) :290–298.

Lazure, P. and Dumas, F. (2008). An external-internal mode coupling for a 3D hydrodynamical model for applications at regional scale (MARS). *Advances in water resources*, 31(2) :233–250.

Lebarbier, É. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4) :717–736.

Leblond, E., Daures, F., Leonardi, S., Demaneche, S., Merrien, C., Berthou, P., Rostiaux, E., Macher, C., Lespagnol, P., Le Grand, C., *et al.* (2014a). Synthèse des flottilles de pêche 2012. flotte de mer du nord-manche-atlantique. flotte de méditerranée.

Leblond, E., Daures, F., Leonardi, S., Demaneche, S., Merrien, C., Berthou, P., Rostiaux, E., Macher, C., Lespagnol, P., Le Grand, C., and Le Blond, S. (2014b). Synthèse des flottilles de pêche 2012. Flotte de Mer du Nord - Manche - Atlantique. Flotte de Méditerranée. http ://archimer.ifremer.fr/doc/00248/35971/.

Leblond, E., Lazure, P., Laurans, M., Rioual, C., Woerther, P., Quemener, L., and Berthou, P. (2010). The RECOPESCA project : a new example of participative approach to collect fisheries and in situ environmental data. *CORIOLIS Quarterly Newsletter*, (37) :40–48.

Lecornu, F. and De Roeck, Y.-H. (2009). Previmer - observations et prévisions côtières. *La Houille Blanche*, (1) :60–63.

Lehuta, S., Petitgas, P., Mahévas, S., Huret, M., Vermard, Y., Uriarte, A., and Record, N. R. (2013). Selection and validation of a complex fishery model using an uncertainty hierarchy. *Fisheries Research*, 143(0) :57 – 66.

Liu, H. and Heynderickx, I. (2009). Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3097–3100. IEEE.

Long, J. A. and Nelson, T. A. (2013). A review of quantitative methods for movement data. *International Journal of Geographical Information Science*, 27(2) :292–318.

Long, J. A., Nelson, T. A., Webb, S. L., and Gee, K. L. (2014). A critical examination of indices of dynamic interaction for wildlife telemetry studies. *Journal of Animal Ecology*, 83(5) :1216–1233.

Madon, B. and Hingrat, Y. (2014). Deciphering behavioral changes in animal movement with a 'multiple change point algorithm-classification tree'framework. *Frontiers in Ecology and Evolution*, 2 :30.

Marchal, P., Bartelings, H., Bastardie, F., Batsleer, J., Delaney, A., Girardin, R., Gloaguen, P., Hamon, K., Hoefnagel, E., Jouanneau, C., Mahevas, S., Nielsen, R., Piwowarczyk, J., Poos, J.-J., Schulze, T., Rivot, E., Simons, S., Tidd, A., Vermard, Y., and Woillez, M. (2014). Mechanisms of change in human behaviour. `http://archimer.ifremer.fr/doc/00223/33377/`.

Martin, P., Muntadas, A., de Juan, S., Sanchez, P., and Demestre, M. (2014). Performance of a northwestern mediterranean bottom trawl fleet : How the integration of landings and VMS data can contribute to the implementation of ecosystem-based fisheries management. *Marine Policy*, 43 :112–121.

Matheron, G. (1978). *Estimer et choisir : essai sur la pratique des probabilités*. Les Cahiers du Centre de morphologie mathématique de Fontainebleau. Ecole nationale supérieure des mines de Paris.

McClintock, B. T., Johnson, D. S., Hooten, M. B., Hoef, J. M. V., and Morales, J. M. (2014). When to be discrete : the importance of time formulation in understanding animal movement. *Movement Ecology*, 2(1) :21.

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., and Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs*, 82(3) :335–349.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*, volume 274. Wiley, New York.

Merrill, E., Sand, H., Zimmermann, B., McPhee, H., Webb, N., Hebblewhite, M., Wabakken, P., and Frair, J. L. (2010). Building a mechanistic understanding of predation with gps-based movement data. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1550) :2279–2288.

Mills, C. M., Townsend, S. E., Jennings, S., Eastwood, P. D., and Houghton, C. A. (2007). Estimating high resolution trawl fishing effort from satellite-based vessel monitoring system data. *Ices Journal of Marine Science*, 64(2) :248–255.

Moorcroft, P. R. and Lewis, M. A. (2013). *Mechanistic home range analysis.(MPB-43)*. Princeton University Press.

Moorcroft, P. R., Lewis, M. A., and Crabtree, R. L. (2006). Mechanistic home range models capture spatial patterns and dynamics of coyote territories in yellowstone. *Proceedings of the Royal Society B-Biological Sciences*, 273(1594) :1651–1659.

Morales, J. and Ellner, S. (2002). Scaling up animal movements in heterogeneous landscapes : The importance of behavior. *Ecology*, 83(8) :2240–2247.

Morales, J., Haydon, D., Frair, J., Holsiner, K., and Fryxell, J. (2004). Extracting more out of relocation data : Building movement models as mixtures of random walks. *Ecology*, 85(9) :2436–2445.

Morales, J. M., Moorcroft, P. R., Matthiopoulos, J., Frair, J. L., Kie, J. G., Powell, R. A., Merrill, E. H., and Haydon, D. T. (2010). Building the bridge between animal movement and population dynamics. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1550) :2289–2301.

Mörters, P. and Peres, Y. (2010). *Brownian motion*, volume 30. Cambridge University Press.

Natale, F., Gibin, M., Alessandrini, A., Vespe, M., and Paulrud, A. (2015). Mapping fishing effort through ais data. *Plos One*, 10(6) :e0130746.

Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., and Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences og the United States of America*, 105(49) :19052–19059.

Newton, I. (1756). *Principes mathématiques de la philosophie naturelle*. Volume 2 de Principes mathématiques de la philosophie naturelle. Chez Desaint & Saillant.

Nielson, M., R., Sawyer, H., and McDonald, T. L. (2013). *BBMM : Brownian bridge movement model*. R package version 3.0.

Norris, J. R. (1998a). *Markov chains*. Number 2008. Cambridge university press.

Norris, J. R. (1998b). *Markov chains*. Number 2008. Cambridge university press.

Øksendal, B. (2003). *Stochastic differential equations*. Springer.

Owen-Smith, N., Fryxell, J. M., and Merrill, E. H. (2010). Foraging theory upscaled : the behavioural ecology of herbivore movement. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1550) :2267–2278.

Palmer, M. C. and Wigley, S. E. (2009). Using positional data from vessel monitoring systems to validate the logbook-reported area fished and the stock allocation of commercial fisheries landings. *North American Journal of Fisheries Management*, 29(4) :928–942.

Patterson, T. A., Basson, M., Bravington, M. V., and Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden markov models. *Journal of Animal Ecology*, 78(6) :1113–1123.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State-space models of individual animal movement. *Trends in Ecology and Evolution*, 23(2) :87–94.

Pearson, K. The problem of the random walk. *Nature*, 72 :294.

Pedersen, A. (1995). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, 1 :257–279.

Pedersen, M. W., Patterson, T. A., Thygesen, U. H., and Madsen, H. (2011). Estimating animal behavior and residency from movement data. *Oikos*, 120(9) :1281–1290.

Peel, D. and Good, N. M. (2011). A hidden markov model approach for determining vessel activity from vessel monitoring system data. *Canadian Journal of Fisheries and Aquatic Sciences*, 68(7) :1252–1264.

Pelletier, D. and Ferraris, J. (2000). A multivariate approach for defining fishing tactics from commercial catch and effort data. *Canadian Journal of Fisgeries and Aquatic Sciences*, 57(1) :51–65.

Pelletier, D., Mahevas, S., drouineau, H., Vermard, Y., Thebaud, O., Guyader, O., and Poussind, B. (2009). Evaluation of the bioeconomic sustainability of multi-species multi-fleet fisheries under a wide range of policy options using isis-fish. *Ecological Modelling*, 220(7) :1013–1033.

Petitgas, P. (1996). Geostatistics and their applications to fisheries survey data. In *Computers in fisheries research*, pages 113–142. Springer.

Picard, F. (2005). *Process segmentation/clustering. Application to the analysis of CGH microarray data*. PhD thesis, Université de Paris XI Orsay.

Picard, F., Robin, S., Lebarbier, E., and Daudin, J.-J. (2007). A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63(3) :758–766.

Pinto, C. and Spezia, L. (2015). Markov switching autoregressive models for interpreting vertical movement data with application to an endangered marine apex predator. *Methods in Ecology and Evolution*, pages n/a–n/a.

Poos, J.-J. and Rijnsdorp, A. D. (2007). The dynamics of small-scale patchiness of plaice and sole as reflected in the catch rates of the dutch beam trawl fleet and its implications for the fleet dynamics. *Journal of Sea Research*, 58(1) :100–112.

Powell, R. (2000). *Research technologies in animal ecology–controversies and consequences*, chapter Animal home ranges and territories and home ranges estimators, pages 65–110. Columbia University press, New York.

Preisler, H. K., Ager, A. A., Johnson, B. K., and Kie, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics*, 15(7) :643–657.

Preisler, H. K., Ager, A. A., and Wisdom, M. J. (2013). Analyzing animal movement patterns using potential functions. *Ecosphere*, 4(3).

Ptolemée (1998). *Ptolemy's Almagest ; transl. and annoted by G. J. Toomer*. Princeton (N.J.) : Princeton university press.

R Core Team (2014). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Rijnsdorp, A. D., Poos, J. J., and Quirijns, F. J. (2011). Spatial dimension and exploitation dynamics of local fishing grounds by fishers targeting several flatfish species. *Canadian Journal of Fisheries and Aquatic Sciences*, 68(6) :1064–1076.

Roberts, G. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88 :603–621.

Roberts, S., Guilford, T., Rezek, I., and Biro, D. (2004). Positional entropy during pigeon homing : application of bayesian latent state modelling. *Journal of Theoretical Biology*, 227(1) :39–50.

Russo, T., D'Andrea, L., Parisi, A., and Cataudella, S. (2014a). VMSbase : An r-package for VMS and logbook data management and analysis in fisheries ecology. *Plos One*, 9(6).

Russo, T., Parisi, A., and Cataudella, S. (2011a). New insights in interpolating fishing tracks from VMS data for different metiers. *Fisheries Research*, 108(1) :184–194.

Russo, T., Parisi, A., Garofalo, G., Gristina, M., Cataudella, S., and Fiorentino, F. (2014b). SMART : A spatially explicit bio-economic model for assessing and managing demersal fisheries, with an application to italian trawlers in the strait of sicily. *Plos One*, 9(1).

Russo, T., Parisi, A., Prorgi, M., Boccoli, F., Cignini, I., Tordoni, M., and Cataudella, S. (2011b). When behaviour reveals activity : Assigning fishing effort to m tiers based on VMS data using artificial neural networks. *Fisheries Research*, 111(1-2) :53–64.

Rutishauser, U. and Koch, C. (2007). Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision*, 7(6) :5.

Schick, R. S., Loarie, S. R., Colchero, F., Best, B. D., Boustany, A., Conde, D. A., Halpin, P. N., Joppa, L. N., McClellan, C. M., and Clark, J. S. (2008). Understanding movement data and movement processes : current and emerging directions. *Ecology Letters*, 11(12) :1338–1350.

Schoener, T. W. (1971). Theory of feeding strategies. *Annual review of ecology and systematics*, pages 369–404.

Seather, S. (1992). The performance of six popular bandwidth selection methods on some real data sets. *COMPUTATIONAL STATISTICS QUARTERLY*, 7 :225–225.

Sermaidis, G. (2010). *Likelihood based inference for discretely observed diffusions*. PhD thesis, University of Warwick.

Shumway, R. and Stoffer, D. (2000). *Time Series Analysis and Its Applications*. Springer Texts In Statistics. Springer-Verlag GmbH.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Skaar, K. L., Jorgensen, T., Ulvestad, B. K. H., and Engas, A. (2011). Accuracy of VMS data from norwegian demersal stern trawlers for estimating trawled areas in the barents sea. *ICES Journal of Marine Science*, 68(8) :1615–1620.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika*, pages 196–218.

Smouse, P. E., Focardi, S., Moorcroft, P. R., Kie, J. G., Forester, J. D., and Morales, J. M. (2010). Stochastic modelling of animal movement. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1550) :2201–2211.

Sorensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time : a survey. *Internat. Statist. Rev.*, 72 :337–354.

Särkkä, S. (2012). Applied stochastic differential equations. `http://users.aalto.fi/~ssarkka/course_s2012/pdf/sde_course_booklet_2012.pdf`.

Stacey, P. C., Walker, S., and Underwood, J. D. (2005). Face processing and familiarity : Evidence from eye-movement data. *British Journal of Psychology*, 96(4) :407–422.

Swain, D. P. and Wade, E. J. (2003). Spatial distribution of catch and effort in a fishery for snow crab (chionoecetes opilio) : tests of predictions of the ideal free distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, 60(8) :897–909.

Tang, W. and Bennett, D. A. (2010). Agent-based modeling of animal movement : A review. *Geography Compass*, 4(7) :682–700.

Turchin, P. (1998). *Quantitative Analysis of Movement : Measuring and Modeling Population Redistribution in Animals and Plants*. Weimar and Now ; 13. Sinauer.

Uchida, M. and Yoshida, N. (2005). AIC for ergodic diffusion processes from discrete observations. *preprint MHF*, 12.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Phys. Rev.*, 36 :823–841.

Van Nieuland, S., Baetens, J. M., De Meyer, H., and De Baets, B. (2015). An analytical description of the time-integrated brownian bridge. *Computational and Applied Mathematics*, pages 1–19.

Vermard, Y., Marchal, P., Mahevas, S., and Thebaud, O. (2008). A dynamic model of the Bay of Biscay pelagic fleet simulating fishing trip choice : the response to the closure of the european anchovy (*Engraulis encrasicolus*) fishery in 2005. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(11) :2444–2453.

Vermard, Y., Rivot, E., Mahevas, S., Marchal, P., and Gascuel, D. (2010). Identifying fishing trip behaviour and estimating fishing effort from VMS data using bayesian hidden markov models. *Ecological Modelling*, 221(15) :1757–1769.

Walker, E. and Bez, N. (2010). A pioneer validation of a state-space model of vessel trajectories (VMS) with observers' data. *Ecological Modelling*, 221(17) :2008–2017.

Walker, E., Rivoirard, J., Gaspar, P., and Bez, N. (2015). From forager tracks to prey distributions : an application to tuna vessel monitoring systems (VMS). *Ecological Applications*, 25(3) :826–833.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411) :699–704.

Worton, B. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70(1) :164–168.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series : An Introduction Using R*. CRC Press.

Annexe A

Suppléments à l'article "An autoregressive model to describe fishing vessel activity"

A.1 Getting the interpolated path from the velocity process

Let $X_0, \ldots, X_t$ be the observed positions. The velocity process as it is described here, is equivalent to the interpolated path. Indeed, given the velocity process, the first position X_0 and the first "heading" θ_1 of an object, the interpolated path can be fully recovered. Indeed, using, as in the text,

$$V_j^p = V_j \cos(\psi_j)$$

$$V_j^r = V_j \sin(\psi_j)$$

let's pose $z_j = V_j^p + iV_j^r$, then $V_j = |z_j|, \psi_j = \arg(z_j), j \geq 2$ and $\psi_1 = \theta_1$. Then observed positions X_k are obtained using

$$X_k = X_0 + \sum_{q=1}^{k} V_q \begin{pmatrix} \cos(\sum_{j=1}^{q} \psi_j) \\ \sin(\sum_{j=1}^{q} \psi_j) \end{pmatrix}$$

Then, modelling velocity is equivalent to a certain description of the path.

A.2 Implementing the Baum Welch algorithm

A.2.1 Notations

Here we define a general Hidden Markov model. The hidden state sequence is a first order Markov chain.

- N hidden states $1, \ldots, N$
- An unobserved hidden state sequence $S_1^T = S_1 \ldots, S_T$, $S_t \in \{1, \ldots, N\}$
- A homogeneous transition $N \times N$ matrix A, $A_{ij} = \mathbb{P}(S_{t+1} = j | S_t = i)$

- T observations $(V_1, \ldots, V_T)$ where $V_t = \begin{pmatrix} V_t^p \\ V_t^r \end{pmatrix}$
- Density of observation according to a state, depending on parameters. $b_i(V_t) = f(V_t|S_t = i, \Theta_i)$
- Initial distribution for states $\pi = \{\pi_i\}_{1 \leq i \leq N}$, $\pi_i = \mathbb{P}(S_1 = i)$
- We note $\Theta = (A, \{\Theta_i\}_{1 \leq i \leq N}, \pi)$ the set of all parameters
- A sequence from l to t is noted $V_l^t = V_l \ldots V_t$

As in Rabiner (1989), 3 problems are defined

1. Given Θ and $V = V_1^T$ an observed sequence, how to calculate the likelihood of the observation ?
2. How to find Θ which maximizes the likelihood ?
3. Given Θ and V, how do we choose the optimal hidden state sequence $S = S_1^T$?

A.2.2 Answer to problem 1 : the *forward-backward* algorithm.

This part is general and (almost) exactly the same as Rabiner (1989), although more detailed some times. It's implementation is necessary to perform the EM algorithm which will give an answer to problem 2. We suppose here that the set of parameters Θ is known. This allows to compute different quantities over the hidden state sequence.

Forward procedure

The forward procedure allows, knowing Θ, to calculate the likelihood of the observed sequence. We define the following density

$$\alpha_t(i) = p(V_1^t, S_t = i|\Theta)$$

One can notice that, using Bayes' formula,

$$\begin{aligned} p(V_1^t, S_t = i|\Theta) &= p(V_t|\Theta,\ V_1^{t-1},\ S_t = i) \times p(V_1^{t-1},\ S_t = i|\Theta) \\ &= b_i(V_t) \times \sum_{j=1}^{N} [p(V_1^{t-1}, S_{t-1} = j|\Theta) \times \mathbb{P}(S_{t-1} = j|S_t = i)] \\ &= b_i(V_t) \times \sum_{j=1}^{N} \alpha_{t-1}(j) A_{ji} \end{aligned}$$

Thus, the following forward procedure is natural to calculate sequence of α :

1. Initialisation :
$$\alpha_1(i) = \pi_i b_i(V_1)$$
2. Induction :
$$\alpha_{t+1}(i) = \big[\sum_{j=1}^{N} \alpha_t(j) A_{ji}\big] b_i(V_{t+1})$$

Knowing α_T allows to calculate the likelihood of the observation, indeed, we have by definition :

$$p(V|\Theta) = \sum_{i=1}^{N} \alpha_T(i). \tag{A.1}$$

Thus, the answer to problem 1 is then known. However, in order to solve problem 2 later, let's make a last effort and compute another algorithm.

***Backward* procedure**

We define the following quantity :

$$\beta_t(i) = p(V_{t+1}^T|V_t,\ S_t = i, \Theta)$$

Then, using Baye's formula, it natural to calculate it by induction :

1. Initialisation :
$$\beta_T(i) = 1$$
2. Induction :
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(V_{t+1}) \beta_{t+1}(j)$$

Both these algorithms are easy to compute in **R** for instance. They are necessary in order to solve problem 2. Actually, two more probabilities are essential for this, the first one is the probability of being in a given state knowing the observed sequence.

$$\gamma_t(i) = \mathbb{P}(S_t = i|V, \Theta)$$

It occurs that this probability is directly computable from αs and βs : x_{10}

$$\begin{aligned}
\gamma_t(i) &= \frac{p(S_t = i,\ V|\Theta)}{p(V|\Theta)} \\
&= \frac{p(S_t = i,\ V_1^t|V_{t+1}^T,\ \Theta) \times p(V_{t+1}^T|\Theta)}{p(V|\Theta)} \\
&= \frac{p(S_t = i,\ V_1^t|\Theta) \times p(V_{t+1}^T|S_t = i,\ V_1^t,\ \Theta) \times p(V_{t+1}^T|\Theta)}{p(V|\Theta) \times p(V_{t+1}^T|\Theta)} \\
&= \frac{p(S_t = i,\ V_1^t|\Theta) \times p(V_{t+1}^T|S_t = i,\ \Theta)}{p(V|\Theta)} && \text{par propriété de Markov} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} p(V|V_t,\ S_t = i,\ \Theta) \times \mathbb{P}(S_t = i|\Theta)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} p(V_1^t|S_t = i,\ \Theta) \times p(V_{t+1}^T|S_t = i,\ \Theta) \times \mathbb{P}(S_t = i|\Theta)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} p(V_1^t,\ S_t = i|\Theta) \times p(V_{t+1}^T|S_t = i,\ \Theta)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}
\end{aligned}$$

The last probability to compute is

$$\xi_t(i,j) = \mathbb{P}(S_t = i,\ S_{t+1} = j | V,\ \Theta).$$

which can be compute directly from :

$$\begin{aligned}
\xi_t(i,j) &= \frac{p(S_t = i,\ S_{t+1} = j,\ V|\Theta)}{\sum_{i,j}^N p(S_t = i,\ S_{t+1} = j,\ V|\Theta)} \\
&= \frac{p(S_{t+1} = j,\ V_{t+1}^T | S_t = i,\ V_1^t,\ \Theta) \times p(S_t = i,\ V_1^t|\Theta)}{\sum_{i,j}^N p(S_t = i,\ S_{t+1} = j,\ V|\Theta)} \\
&= \frac{p(V_{t+2}^T | S_{t+1} = j,\ S_t = i,\ V_1^{t+1},\ \Theta) \times p(S_{t+1} = j,\ V_{t+1} | S_t = i,\ V_1^t,\ \Theta) \times p(S_t = i,\ V_1^t|\Theta)}{\sum_{i,j}^N p(S_t = i,\ S_{t+1} = j,\ V|\Theta)} \\
&= \frac{\beta_{t+1}(j) p(V_{t+1} | S_{t+1} = j,\ S_t = i,\ V_1^t,\ \Theta) \times p(S_{t+1} = j | S_t = S_i,\ V_1^t,\ \Theta)\alpha_t(i)}{\sum_{i,j}^N p(S_t = i,\ S_{t+1} = j,\ V|\Theta)} \\
&= \frac{\beta_{t+1}(j) b_j(V_{t+1}) a_{ij} \alpha_t(i)}{\sum_{i,j}^N \beta_{t+1}(j) b_j(V_{t+1}) a_{ij} \alpha_t(i)}
\end{aligned}$$

Let's see now how these quantities are useful to estimate likelihood.

A.2.3 An answer to problem 2 : The Baum Welch algortihm

The Baum Welch algorithm is a variant of EM, which is described in McLachlan and Krishnan (1997) for instance. We present here the equations for the model presented in the paper. The EM algorithm is iterative. At each step k, we suppose known the real set of parameters, noted $\Theta^{(k)}$.
E (Expectation) step : Compute

$$Q(\Theta, \Theta^{(k)}) = \mathbb{E}(\log(p(V, S; \Theta)|V = V_{obs},\ \Theta^{(k)})$$

that is the expectation of the complete data loglikelihood conditionnaly to observed data and true set of parameters $\Theta^{(k)}$.
M (Maximization) step : find $\Theta^{(k+1)} = \text{argmax}_\Theta Q(\Theta, \Theta^{(k)})$. Then (Baum *et al.* (1970))

Proposition A.2.1.

$$Q(\Theta^{(k+1)}, \Theta^{(k)}) \geq Q(\Theta^{(k)}, \Theta^{(k)}) \Rightarrow L(\Theta^{(k+1)}) \geq L(\Theta^{(k)})$$

where L is the log likelihood of observed data.

We present here equations for E step and M step in our case

E step

Using the whole sequence, we have

$$\begin{aligned} p(V,S;\Theta) =& \mathbb{P}(S_1 = s_1|\Theta)p(V_1|S_1 = s_1,\Theta) \\ & \times \prod_{t=2}^{T} \mathbb{P}(S_t = s_t|S_{t-1} = s_{t-1},\Theta)p(V_t|S_t = s_t,\Theta) \\ =& \pi_{s_1} b_{s_1}(V_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} b_{s_t}(V_t) \end{aligned}$$

which implies

$$\begin{aligned} Q(\Theta,\Theta^{(k)}) = & \sum_{S=s_1,\dots s_T} \mathbb{P}(S|\ V = V_{obs},\Theta^{(k)})\log(\pi_{s_1}) \\ & + \sum_{S=s_1,\dots s_T} \mathbb{P}(S|\ V = V_{obs},\Theta^{(k)}) \sum_{t=2}^{T} \log(a_{s_{t-1}s_t}) \\ & + \sum_{S=s_1,\dots s_T} \mathbb{P}(S|\ V = V_{obs},\Theta^{(k)}) \sum_{t=1}^{T} \log(b_{s_t}(V_t)) \end{aligned}$$

This can be simplify in

$$Q(\Theta,\Theta^{(k)}) = \sum_j p(S_1 = j|\ V = V_{obs},\Theta^{(k)})\log\pi_j \tag{A.2}$$

$$+ \sum_{i,j=1} \Big(\sum_{t=2}^{T} \mathbb{P}(S_{t-1} = i, S_t = j|\ V = V_{obs},\Theta^{(k)}\Big)\log(a_{ij}) \tag{A.3}$$

$$+ \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{P}(S_t = i|\ V = V_{obs},\Theta^{(k)})\log(b_i(V_t)) \tag{A.4}$$

With the notations of forward backward algorithm :

- $\xi_t(i,j) := \mathbb{P}(S_{t-1} = i, S_t = j|\ V = V_{obs},\Theta^{(k)}\big)$
- $\gamma_t(i) := \mathbb{P}(S_t = i|\ V = V_{obs},\Theta^{(k)})$
- $S_l^m := s_l \dots s_m$.

Lines 2 and 3 are deduced from

$$\begin{aligned} \sum_{S_1^T} \mathbb{P}(S_1^T|\ V_{obs},\Theta^{(k)}) \sum_{t=2}^{T} \log(a_{s_{t-1}s_t}) &= \sum_{i,j=1}^{N} \sum_{t=2}^{T} \sum_{S_1^{t-2},S_{t+1}^T} \mathbb{P}(S_1^{t-2},S_{t+1}^T|S_{t-1} = i,\ S_t = j,\ V_{obs},\Theta^{(k)})\xi_t(i,j)\log(a_{ij}) \\ &= \sum_{i,j=1}^{N} \sum_{t=2}^{T} \xi_t(i,j)\log(a_{ij}) \underbrace{\sum_{S_1^{t-2},S_{t+1}^T} \mathbb{P}(S_1^{t-2},S_{t+1}^T|S_{t-1} = i,\ S_t = j,\ V_{obs},\Theta^{(k)})}_{=1} \end{aligned}$$

$$\begin{aligned} \sum_{S_1^T} \mathbb{P}(S_1^T|\ V_{obs},\Theta^{(k)}) \sum_{t=1}^{T} \log(b_{s_t}(V_t)) &= \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{S_1^{t-1},S_{t+1}^T} \mathbb{P}(S_1^{t-2},S_{t+1}^T|S_{t-1} = i,\ S_t = j,\ V_{obs},\Theta^{(k)})\gamma_t(i)\log(b_i(V_t)) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i)\log(b_i(V_t)) \underbrace{\sum_{S_1^{t-1},S_{t+1}^T} \mathbb{P}(S_1^{t-2},S_{t+1}^T|S_{t-1} = i,\ S_t = j,\ V_{obs},\Theta^{(k)})}_{=1} \end{aligned}$$

After E step, the M step is performed

Étape Maximization

Equations (A.2) and (A.3) have the following MLE

$$\hat{\pi}_i = \gamma_1(i) \tag{A.5}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{A.6}$$

Equation A.6 reflects the relation

$$\frac{\#\text{ expected transtitions from i to j}}{\#\text{ expected transitions from i}}.$$

We then have to maximize (A.4)

$$\sum_{i=1}^{N}\sum_{t=1}^{T} \gamma_t(i)\log(b_i(V_t)) \tag{A.7}$$

As we supposed $V^p \perp V^r$ we write the solutions for an univariate process V. For each state i, one have to maximize over η_i, μ_i and σ_i^2

$$\begin{aligned} l(\eta_i, \mu_i, \sigma_i^2) &:= \sum_{t=2}^{T} \gamma_t(i)\log(b_i(V_t)) \\ &= -n\log(2\pi) - \frac{1}{2}\sum_{t=2}^{T} \gamma_t(i)\log(\sigma_i^2) - \frac{1}{2\sigma_i^2}\sum_{t=2}^{T} \gamma_t(i)(V_t - \eta_i - \mu_i V_{t-1})^2 \end{aligned}$$

MLE $\hat{\eta}_i$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ satisfy

$$\begin{cases} \dfrac{\delta\ l(\eta_i, \mu_i, \sigma_i^2)}{\delta \eta_i}(\hat{\eta}_i, \hat{\mu}_i, \hat{\sigma}_i^2) = 0 \\ \dfrac{\delta\ l(\eta_i, \mu_i, \sigma_i^2)}{\delta \mu_i}(\hat{\eta}_i, \hat{\mu}_i, \hat{\sigma}_i^2) = 0 \\ \dfrac{\delta\ l(\eta_i, \mu_i, \sigma_i^2)}{\delta \sigma_i^2}(\hat{\eta}_i, \hat{\mu}_i, \hat{\sigma}_i^2) = 0 \end{cases} \tag{A.8}$$

Given that we have

$$\begin{aligned} \frac{\delta\ l(\eta_i, \mu_i, \sigma_i^2)}{\delta \eta_i}(\eta_i, \mu_i, \sigma_i^2) &= \frac{1}{\sigma_i^2}\sum_{t=2}^{T} \gamma_t(i)(V_t - \eta_i - \mu_i V_{t-1}) \\ \frac{\delta\ l(\eta_i, \mu_i, \sigma_i^2)}{\delta \mu_i}(\eta_i, \mu_i, \sigma_i^2) &= \frac{1}{\sigma_i^2}\sum_{t=2}^{T} \gamma_t(i)V_{t-1}(V_t - \eta_i - \mu_i V_{t-1}) \\ \frac{\delta\ l(\eta_i, \mu_i, \sigma_i^2)}{\delta \sigma_i^2}(\eta_i, \mu_i, \sigma_i^2) &= -\frac{1}{2\sigma_i^2}\sum_{t=2}^{T} \gamma_t(i) + \frac{1}{2\sigma_i^2}\sum_{t=2}^{T} \gamma_t(i)(V_t - \eta_i - \mu_i V_{t-1})^2 \end{aligned}$$

Thus, (A.8) is equivalent to

$$\begin{cases} \hat{\eta}_i \sum_{t=2}^{T} \gamma_t(i) + \hat{\mu}_i \sum_{t=2}^{T} \gamma_t(i) V_{t-1} = \sum_{t=2}^{T} \gamma_t(i) V_t \\ \hat{\eta}_i \sum_{t=2}^{T} \gamma_t(i) V_{t-1} + \hat{\mu}_i \sum_{t=2}^{T} \gamma_t(i) V_{t-1}^2 = \sum_{t=2}^{T} \gamma_t(i) V_t V_{t-1} \\ \hat{\sigma}_i^2 = \frac{\sum_{t=2}^{T} \gamma_t(i)(V_t - \hat{\eta}_i - \hat{\mu}_i V_{t-1})^2}{\sum_{t=2}^{T} \gamma_t(i)} \end{cases} \tag{A.9}$$

Which is again equivalent to

$$\begin{cases} \hat{\eta}_i = \frac{\sum_{t=2}^{T} \gamma_t(i)(V_t - \hat{\mu}_i V_{t-1})}{\sum_{t=2}^{T} \gamma_t(i)} \\ \hat{\mu}_i = \frac{\sum_{t=2}^{T} \gamma_t(i) \times \sum_{t=2}^{T} \gamma_t(i) V_t V_{t-1} - \sum_{t=2}^{T} \gamma_t(i) V_{t-1} \times \sum_{t=2}^{T} \gamma_t(i) V_t}{\sum_{t=2}^{T} \gamma_t(i) \times \sum_{t=2}^{T} \gamma_t(i) V_{t-1}^2 - (\sum_{t=2}^{T} \gamma_t(i) V_{t-1})^2} \\ \hat{\sigma}_i^2 = \frac{\sum_{t=2}^{T} \gamma_t(i)(V_t - \hat{\eta}_i - \hat{\mu}_i V_{t-1})^2}{\sum_{t=2}^{T} \gamma_t(i)} \end{cases} \tag{A.10}$$

A.2.4 An answer to problem 3, the Viterbi algorithm

Viterbi algorithm consists in computing

$$\delta_t(i) = \max_{S_1,\ldots,S_{t-1}} p(S_1 \ldots S_t = i,\ V|\Theta)$$

Again, by induction, one can notice that

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(V_{t+1}).$$

The Viterbi algorithm therefore consists in :

Initialization

$$\delta_1(i) = \pi_i b_i(V_1)$$
$$\psi_1(i) = 0.$$

Iteration

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(V_t)$$
$$\psi_t(j) = \operatorname{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]$$

Getting the best path

$$\hat{S}_T = \operatorname{argmax}_{1 \le i \le N} [\delta_T(i)]$$
$$\hat{S}_t = \psi_{t+1}(\hat{S}_{t+1})$$

A.3 Results on simulated scenarios for the V_r process

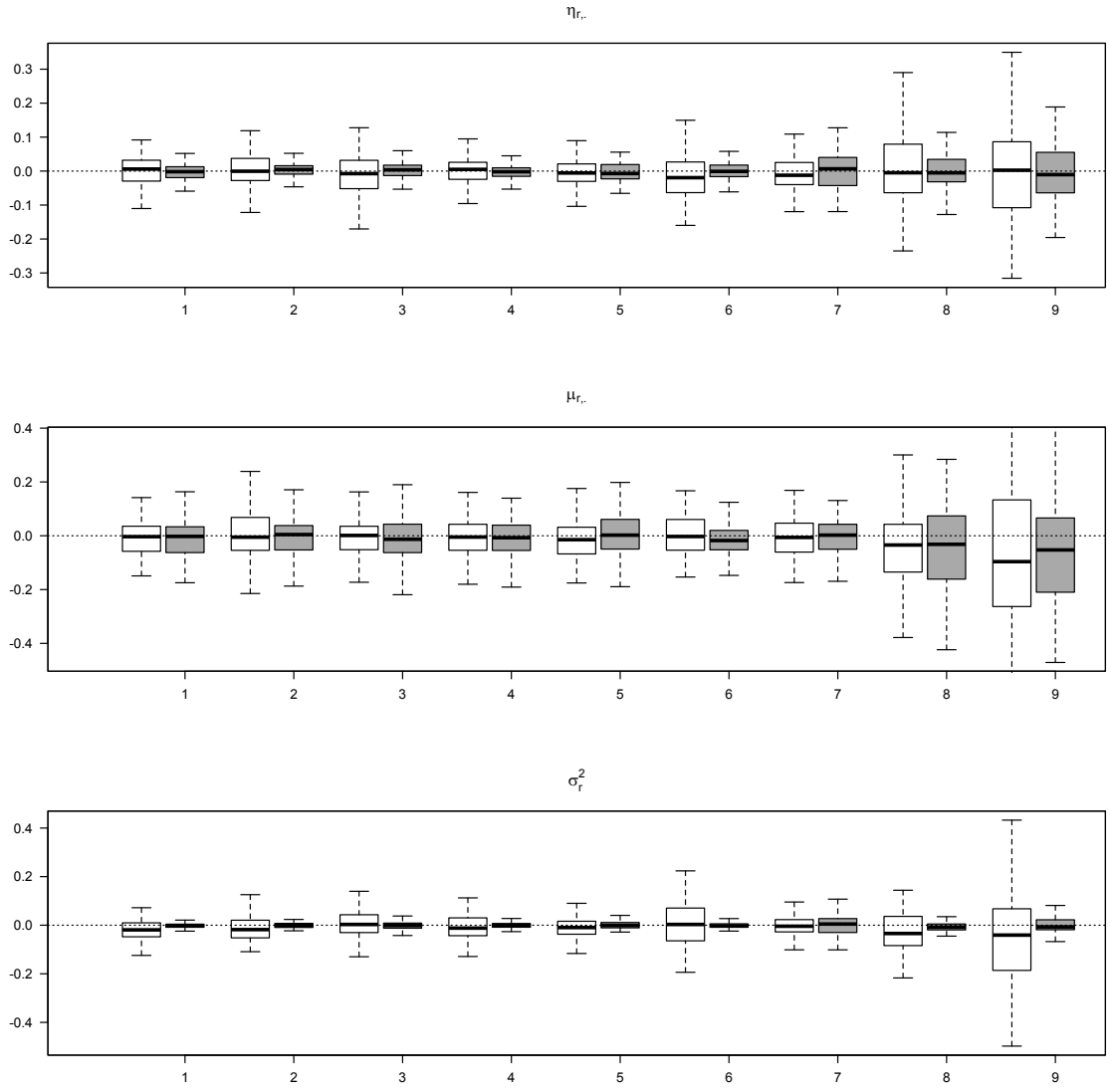


FIGURE *A.1 – Estimation errors (estimated value minus the true value of each parameter, on the y axis) obtained for the 9 simulation scenarios (x axis) presented in Table 2.2. Box plots represent the variability between the 100 simulations for each scenario. Only estimation errors for process V^r are presented, white and grey box plots are for parameters estimates in steaming and fishing respectively. The whiskers represent here at most 1.5 times the interquartile range. Outliers are not plotted.*

A.4 Equivalence between the MLE of an AR process and the MLE of a regular sampled Ornstein Ulhenbeck process

The goal is to show that there is a diffeomorphism of class C^1 between the MLE of a positively auto correlated AR process and the MLE of a regular sampled OU process (with the same time lag as the AR process).
The OU process sampled at regular discrete times is defined with the following equation :

$$V_t = a(1 - e^{-b}) + e^{-b}V_{t-1} + c\sqrt{\frac{1 - e^{-2b}}{2b}}\epsilon_t \qquad \text{Regular sampled OU process} \qquad (A.11)$$

with $a \in \mathbb{R}, \quad b \in \mathbb{R}_+^*, \quad c \in \mathbb{R}_+^*$.
A first order, positively correlated, AR process, sample with the same time lag is defined as follows :

$$V_t = \alpha + \beta V_{t-1} + \sqrt{\gamma}\epsilon_t \qquad \text{AR process} \qquad (A.12)$$

with $\alpha \in \mathbb{R}, \quad \beta \in]0; 1[, \quad \gamma \in \mathbb{R}_+^*$

Let's $\mathbf{V} = \{V_t\}_{t\in 0\ldots n}$ be the observations, and suppose (for simplicity) V_0 to be a known constant. The log likelihood of the observations with respect to the OU process defined in equation (A.11) is

$$\begin{aligned} l_{OU}(\mathbf{V},\ a,\ b,\ c) &= \sum_{i=1}^{N} \log(p(V_t|V_{t-1}\ ,\ a,\ b,\ c)) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(c^2\frac{1 - e^{-2b}}{2b}) - \frac{1}{2c^2\frac{1-e^{-2b}}{2b}}\sum_{t=1}^{n}(V_t - e^{-b}V_{t-1} - a(1 - e^{-b}))^2 \end{aligned} \qquad (A.13)$$

In a similar way, the log likelihood of the observations with respect to the AR process defined in equation (A.12) is

$$l_{AR}(\mathbf{V},\ \alpha,\ \beta,\ \gamma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\gamma) - \frac{1}{2\gamma}\sum_{t=1}^{n}(V_t - \beta V_{t-1} - \alpha)^2 \qquad (A.14)$$

Let's now define $g_{\mathbf{V}}$

$$\begin{aligned} g_{\mathbf{V}} \quad \mathbb{R}\times]0; 1[\times\mathbb{R}_+^* &\longrightarrow \mathbb{R} \\ (\alpha, \beta, \gamma) &\longrightarrow l_{AR}(\mathbf{V},\ \alpha,\ \beta,\ \gamma) \end{aligned} \qquad (A.15)$$

A.4.1 The MLE of the AR process

The MLE of the AR process defined in (A.12) is the triplet $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ such that $Dg(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = 0$ (where D is the usual differential operator). This triplet is given by (see appendix A.2)

$$\hat{\alpha} = \frac{\sum_{t=1}^{n}(V_t - \hat{\beta}V_{t-1})}{n} \tag{A.16}$$

$$\hat{\beta} = \frac{n\sum_{t=1}^{n} V_t V_{t-1} - \sum_{t=1}^{n} V_t \times \sum_{t=1}^{n} V_{t-1}}{n\sum_{t=1}^{n}(V_{t-1})^2 - (\sum_{t=1}^{n} V_{t-1})^2} \tag{A.17}$$

$$\hat{\gamma} = \frac{\sum_{t=1}^{n}(V_t - \hat{\beta}V_{t-1} - \hat{\alpha})^2}{n} \tag{A.18}$$

A.4.2 Existence of a diffeomorphism of class C^1

Proposition : There exists a diffeomorphism f of class C^1 (which is sufficient here) from $\mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ to $\mathbb{R}\times]0,1[\times\mathbb{R}_+^*$ such that

$$f^{-1}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = (\frac{\hat{\alpha}}{1-\hat{\beta}}; -\log\hat{\beta}; \sqrt{\frac{-2\hat{\gamma}\log\hat{\beta}}{1-\hat{\beta}^2}})^T$$

is the MLE of the regularly sampled OU process defined in equation (A.11).

The f function

Let us define define a function f :

$$\begin{aligned} f: \quad \mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R}_+^* &\longrightarrow \mathbb{R}\times]0,1[\times\mathbb{R}_+^* \\ \begin{pmatrix} a \\ b \\ c \end{pmatrix} &\longrightarrow \begin{pmatrix} a(1-e^{-b}) \\ e^{-b} \\ c^2\frac{1-e^{-2b}}{2b} \end{pmatrix} \end{aligned} \tag{A.19}$$

f is one to one

Suppose $X_1 = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}$, $X_2 = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}$

$$
\begin{aligned}
f(X_1) = f(X_2) \Leftrightarrow & \begin{cases} x_1(1-\mathrm{e}^{-y_1}) = x_2(1-\mathrm{e}^{-y_2}) \\ \mathrm{e}^{-y_1} = \mathrm{e}^{-y_2} \\ z_1^2 \frac{1-\mathrm{e}^{-y_1}}{2y_1} = z_2^2 \frac{1-\mathrm{e}^{-y_2}}{2y_2} \end{cases} \\
\Leftrightarrow & \begin{cases} x_1(1-\mathrm{e}^{-y_1}) = x_2(1-\mathrm{e}^{-y_1}) \\ y_1 = y_2 \\ z_1^2 \frac{1-\mathrm{e}^{-y_1}}{2y_1} = z_2^2 \frac{1-\mathrm{e}^{-y_1}}{2y_1} \end{cases} \\
\Leftrightarrow & \begin{cases} x_1 = x_2 \\ y_1 = y_2 \\ z_1 = z_2 \text{ As they are both positive numbers} \end{cases} \\
\Leftrightarrow & \quad X_1 = X_2
\end{aligned} \tag{A.20}
$$

f is a surjection

Let $X_2 = (x_2; y_2; z_2)^T \in \mathbb{R}\times]0,1[\times\mathbb{R}_+^*$. One can check that if $X_1 = (\frac{x_2}{1-y_2}; -\log(y_2); \sqrt{-2z_2\frac{\log(y_2)}{1-y_2^2}})^T$, then $f(X_1) = X_2$. Then, f is a surjection (and, from above, a bijection), with inverse function

$$
\begin{aligned}
f^{-1}: \quad \mathbb{R}\times]0,1[\times\mathbb{R}_+^* &\longrightarrow \mathbb{R}\times\mathbb{R}_+^*\times\mathbb{R}_+^* \\
\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} &\longrightarrow \begin{pmatrix} \frac{\alpha}{1-\beta} \\ -\log\beta \\ \sqrt{\frac{-2\gamma\log\beta}{1-\beta^2}} \end{pmatrix}
\end{aligned} \tag{A.21}
$$

f is a diffeomorphism

The Jacobian matrix of f is

$$
Df(a,b,c) = \begin{pmatrix} 1-\mathrm{e}^{-b} & 0 & 0 \\ a\mathrm{e}^{-b} & -\mathrm{e}^{-b} & c^2\frac{\mathrm{e}^{-2b}(1+b)-1}{2b^2} \\ 0 & 0 & c\frac{1-\mathrm{e}^{-2b}}{b} \end{pmatrix} \tag{A.22}
$$

The following property then holds :

$$
\forall (a,b,c) \in \mathbb{R}\times\mathbb{R}_+^*\times\mathbb{R}_+^*, \quad \det(Df(a,b,c)) \neq 0
$$

thus, f is a C^1 bijection such that the determinant of the Jacobian is non 0, by definition, it is a diffeomorphism of class C^1 from $\mathbb{R}\times\mathbb{R}_+^*\times\mathbb{R}_+^*$ to $\mathbb{R}\times]0;1[\times\mathbb{R}_+^*$.

A.4.3 The MLE of the OU process

From (A.19) and (A.15) one can notice that

$$\begin{aligned}(g_{\mathbf{V}} \circ f)(a,b,c) &= g_{\mathbf{V}}\left(a(1-e^{-b}); e^{-b}; c^2\frac{1-e^{-2b}}{2b}\right)\\ &= l_{OU}(\mathbf{V},a,b,c)\end{aligned} \tag{A.23}$$

Therefore,

$$Dl_{OU}(\mathbf{V},a,b,c) = Df(a,b,c) \times Dg_{\mathbf{V}}(f(a,b,c))$$

It follows that

$$\begin{aligned} & Dl_{OU}(\mathbf{V},\hat{a},\hat{b},\hat{c}) = 0 \\ \Leftrightarrow \quad & Df(\hat{a},\hat{b},\hat{c}) \times Dg_{\mathbf{V}}(f(\hat{a},\hat{b},\hat{c})) = 0 \end{aligned} \tag{A.24}$$

As f is a diffeomorphism, this is equivalent to

$$\Leftrightarrow \quad Dg_{\mathbf{V}}(f(\hat{a},\hat{b},\hat{c})) = 0 \tag{A.25}$$

$$\Leftrightarrow \quad f(\hat{a},\hat{b},\hat{c}) = (\hat{\alpha},\hat{\beta},\hat{\gamma})^T \tag{A.26}$$

$$\Leftrightarrow \quad (\hat{a},\hat{b},\hat{c})^T = f^{-1}(\hat{\alpha},\hat{\beta},\hat{\gamma}) = (\frac{\hat{\alpha}}{1-\hat{\beta}}, -\log\hat{\beta}, \sqrt{\frac{-2\hat{\gamma}\log\hat{\beta}}{1-\hat{\beta}^2}})^T \tag{A.27}$$

Annexe B

Suppléments à l'article "Is the speed in relation to the water mass a better proxy for fishing activities than speed in relation to the ground"

B.1 Removing currents from speed processes

Here we detail the methods used to derive the speed in relation to the ground from the sequence of recorded GPS positions, and then the method used to derive the speed in relation to the water mass by combining the speed in relation to the ground with surface currents field. The main lines of the method are given below.

Let us denote $(Z_i)_{i=1\ldots(n)}$ the sequence of GPS positions (latitude and longitude) recorded at times $t_1 \ldots t_n$. In the RECOPESCA data set, the time step is 15min. But the method presented here does not depend upon the time step. The positions are converted into a sequence of two dimensional coordinate on a regular grid distance, denoted $(X_i)_{i=1\ldots(n)}$. The sequence of speed in relation to the ground, denoted $(V_i^{raw})_{i=1\ldots(n-1)}$ is then derived from the sequence of positions assuming a linear path between points.

$$V_i^{raw} = \frac{X_{i+1} - X_i}{t_{i+1} - t_i} \tag{B.1}$$

The speed in relation to the ground results from the combination of forces due to the engine and to the surface currents. To remove surface currents component on raw speed process computed in equation (B.1), outputs of physical model MARS 3D are used. MARS 3D is a hydrodynamical model developed by IFREMER (Lazure and Dumas, 2008) operated to derive operational prediction in coastal oceanography (Lecornu and De Roeck, 2009). Surface currents (taking wind in account) are computed on a 4km×4km grid every hour. The model provides the vector of surface current (vector) C_h^P at each point P of the grid and every hour h of the day.The time step and GPS positions $(X_i)_{i=1\ldots(n)}$ do not match with the time step and the grid of MARS3D outputs. Each observed position X_i is recorded at time $t_i = h_i : m_i$ (hour :minute) and is linked to its closest point M in the model's grid. To compute an estimation of the current at position

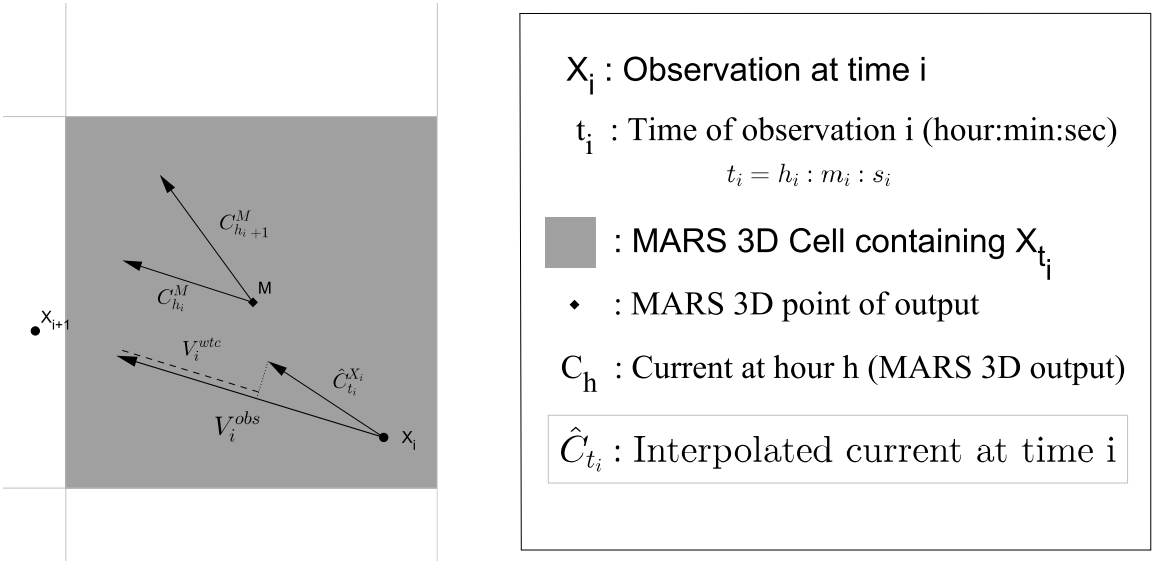


FIGURE *B.1 – Method to remove surface currents from velocity process and to obtain speed without currents.*

X_i and time t_i (noted $\hat{C}^{X_i}_{t_i}$), an interpolation is performed following the equation

$$\hat{C}^{X_i}_{t_i} = (1 - \frac{m_i}{60})C^M_{h_i} + \frac{m_i}{60}C^M_{h_i+1} \tag{B.2}$$

Combining the surface currents $\hat{C}^{X_i}_{t_i}$ in (B.2) with the speed in relation to the ground in (B.1), the speed in relation to the water mass (i.e., after retrieving the surface current), denoted V^{wtc}, is computed as :

$$V_i^{wtc} = V^{raw} - < \frac{V^{raw}}{\parallel V^{raw} \parallel}; \hat{C}^{X_i}_{t_i} > \frac{V^{raw}}{\parallel V^{raw} \parallel} \tag{B.3}$$

where $< \cdot; \cdot >$ is the usual scalar product. A geometric representation of equation (B.3) is given at figure B.1.

Using equation (B.1), the 1 dimension speed process $\parallel V_i^{raw} \parallel$ can be computed.

Annexe C

Suppléments à la partie 3.1.1, Quantifier l'utilisation de l'espace

C.1 Preuve que $\tilde{h}(z) = p(z)$ dans le cas où $\tilde{h}(z,T) = \tilde{h}(z)$

Soit

$$\begin{aligned} f_z : \ ]0;\infty[&\mapsto \mathbb{R}_+ \\ u &\mapsto \tilde{h}(z,u) \end{aligned}$$
$$\text{Avec } h(z,u) = \frac{1}{u}\int_0^u p(z,t)dt$$

Si $\tilde{h}$ ne dépend pas de T, alors, pour z fixé, on a, pout tout T

$$\begin{aligned} & f'_z(T) = 0 \\ \Leftrightarrow \quad & \frac{Tp(z,T) - \int_0^T p(z,t)dt}{T^2} = 0 \\ \Leftrightarrow \quad & Tp(z,T) = \int_0^T p(z,t)dt \end{aligned}$$

Cette dernière égalité est vraie pour tout T. En dérivant chaque membre par rapport à T, on obtient $\frac{dp}{dT} = 0$.

Annexe D

Suppléments à l'article "Stochastic differential equation based on a Gaussian potential field to model fishing vessels trajectories"

D.1 Exact conditional simulation of trajectories

This section details the mechanism to sample Y_u conditionally on (Y_0, Y_T), where $(Y_t)_{0\leq t\leq T}$ is a process solution to (3.28). Using the exact algorithm EA1 of Beskos *et al.* (2006a), this can be done using a rejection sampling algorithm based on Brownian bridges obtained at some random time steps. Let $\mathbb{Q}_\theta^y$ (resp. $\mathbb{W}^y$) be the probability measure induced by $(Y_t)_{0\leq t\leq T}$ (resp. $(W_t)_{0\leq t\leq T}$) on $(\mathsf{C}, \mathcal{C})$ conditioned on hitting y at time T. By (3.27), Girsanov theorem implies that

$$\frac{\mathrm{d}\mathbb{Q}_\theta^y}{\mathrm{d}\mathbb{W}^y}(\omega) \propto \exp\left\{-\frac{1}{2}\int_0^T \left\{\|\alpha_\theta(\omega_s)\|^2 + \Delta H_\theta(\omega_s)\right\}\mathrm{d}s\right\}. \tag{D.1}$$

By lemmas D.2.1 and D.2.2, there exists $\ell_\theta \in \mathbb{R}$ such that for all $x \in \mathbb{R}^2$,

$$\frac{1}{2}\left(\|\alpha_\theta(x)\|^2 + \Delta H_\theta(x)\right) \geq \ell_\theta\,.$$

Then, if ϕ_θ is given by

$$\phi_\theta : x \mapsto \frac{1}{2}\left(\|\alpha_\theta(x)\|^2 + \Delta H_\theta(x)\right) - \ell_\theta, \tag{D.2}$$

the likelihood ratio (D.1) can be written :

$$\frac{\mathrm{d}\mathbb{Q}_\theta^y}{\mathrm{d}\mathbb{W}^y}(\omega) \propto \exp\left\{-\int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right\} \leq 1\,.$$

To draw a path with distribution $\mathbb{Q}_\theta^y$, a path $(\omega_s)_{0\leq s\leq T}$ is sampled from $\mathbb{W}^y$ and accepted with probability p_θ given by

$$p_\theta = \exp\left(-\int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right). \tag{D.3}$$

This acceptance probability cannot be evaluated in practice but this computation can be avoided by accepting a proposed path from $\mathbb{W}^y$ with the realization of an event of probability p_θ.

In Beskos *et al.* (2006b), p_θ is interpreted as the probability of an event associated with an inhomogeneous Poisson process on $[0,T]$ with intensity $\phi_\theta(\omega_s)$. By lemmas D.2.1 and D.2.2, there exists Λ such that

$$\sup_{s\in[0,T]} \phi_\theta(\omega_s) \leq \Lambda_\theta \,. \tag{D.4}$$

By Beskos *et al.*, 2006b, Theorem 1, if Φ is a Poisson process on $[0,T]\times[0,\Lambda_\theta]$ with intensity $\Lambda_\theta T$ and if N_Φ is the number of points of Φ below the graph of $s \mapsto \phi_\theta(\omega_s)$, then

$$\mathbb{P}\left[N_\Phi = 0 | (\omega_s)_{0\leq s\leq T}\right] = p_\theta \,.$$

The mechanism goes as follows. Let M be distributed according to a Poisson random variable with parameter $\Lambda_\theta T$, $\{\tau_1,\ldots,\tau_M\}$ be uniformly distributed on $[0,T]$ and $\{\Phi_1,\ldots,\Phi_M\}$ be uniformly distributed on $[0,\Lambda_\theta]$. Then, a proposed path $(\omega_s)_{0\leq s\leq t}$ from $\mathbb{W}^y$ (sampled using Brownian bridge dynamics) is accepted as a path from $\mathbb{Q}_\theta^y$ if $\mathbb{I}=1$, where

$$\mathbb{I} := \prod_{i=1}^{M} 1_{\phi_\theta(\omega_{\tau_j})<\Lambda_\theta\Phi_j} \,. \tag{D.5}$$

If the proposed path is accepted it can then be filled at time u using brownian bridge dynamics. This procedure is displayed in Algorithm 2.

Algorithm 2 Exact algorithm to sample Y_u conditionally on $(Y_0=x, Y_T=y)$

1: **repeat**
2: Let Λ_θ be such that $\sup_{x\in\mathbb{R}^2}\phi_\theta(x)\leq\Lambda_\theta$.
3: Draw M, a Poisson random variable with parameter $\Lambda_\theta T$.
4: Draw $\{\tau_1,\ldots,\tau_M\}$ uniformly on $[0,T]$ and $\{\Phi_1,\ldots,\Phi_M\}$ uniformly on $[0,\Lambda_\theta]$.
5: Conditionally to $Y_0=x$ and $Y_T=y$, draw a Brownian bridge at times $\{\tau_1,\ldots,\tau_M\}$. The skeleton obtained is noted $\omega_{\tau_1}\ldots\omega_{\tau_M}$
6: Compute
$$\mathbb{I} = \prod_{i=1}^{M} 1_{\phi_\theta(\omega_{\tau_j})<\Lambda_\theta\Phi_j} = 1 \,.$$
7: **until** $\mathbb{I}=1$
8: Find $k\in\{1\ldots M-1\}$ such that $\tau_k\leq u<\tau_{k+1}$
9: Draw Y_u from a Brownian Bridge, starting at ω_{τ_k} and ending at $\omega_{\tau_{k+1}}$.

D.2 Implementation of EA1 for a mixture of Gaussian fields

Lemma D.2.1. *For all θ and all $x\in\mathbb{R}^2$,*

$$0 \leq \|\alpha_\theta(x)\|^2 \leq \bar{\alpha}_\theta \,,$$

where α_θ is given by (3.29) *and, with $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ the eigenvalues of C_k, $1\leq k\leq K$,*

$$\bar{\alpha}_\theta := \mathrm{e}^{-1}\gamma^{-2}\bar{\pi}\sum_{k=1}^{K}\pi_k \max_{1\leq j\leq 2}\lambda_j^{(k)} \quad \textit{and} \quad \bar{\pi} := \sum_{k=1}^{K}\pi_k \,. \tag{D.6}$$

Démonstration. For all θ and all $x \in \mathbb{R}^2$, by convexity of $\|\cdot\|^2$,

$$\begin{aligned}\|\alpha_\theta(x)\|^2 &= \left\| \sum_{k=1}^K \pi_k \varphi_k(\gamma x)\gamma^{-1} C_k(\gamma x - \mu_k) \right\|^2 , \\ &\leq \bar{\pi} \sum_{k=1}^K \pi_k \varphi_k^2(\gamma x) \left\| \gamma^{-1} C_k(\gamma x - \mu_k) \right\|^2 , \\ &\leq \bar{\pi} \sum_{k=1}^K \pi_k \gamma^{-2} \left\| C_k(\gamma x - \mu_k) \right\|^2 \exp\left(-(\gamma x - \mu_k)^T C_k(\gamma x - \mu_k) \right) .\end{aligned}$$

Let Λ_k be defined by $C_k = P_k^{-1}\Lambda_k P_k$ where Λ_k is the diagonal matrix with diagonal given by $\left(\lambda_1^{(k)}, \lambda_2^{(k)}\right)$ and where $P_k P_k^t = I_2$. If $z_k := P_k(\gamma x - \mu_k)$ then,

$$\|\alpha_\theta(x)\|^2 \leq \bar{\pi}\gamma^{-2} \sum_{k=1}^K \pi_k \underbrace{z_k^T \Lambda_k^2 z_k \exp\left(- \left[z_k^T \Lambda_k z_k \right] \right)}_{f(z_k)} ,$$

where we used $\left\| P_k^T z_k \right\|^2 = \|z_k\|^2$ as P_k is orthogonal. The proof is concluded upon noting that for all $x \in \mathbb{R}^2$,

$$f(x) \leq \mathrm{e}^{-1} \max_{1 \leq j \leq 2} \lambda_j^{(k)} .$$

□

Lemma D.2.2. *For all θ and all $x \in \mathbb{R}^2$,*

$$\Delta_\theta^- \leq \Delta H_\theta(x) \leq \Delta_\theta^+ ,$$

where ΔH_θ is given by (3.27) *and, with $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ the eigenvalues of C_k, $1 \leq k \leq K$,*

$$\Delta_\theta^- := -\sum_{k=1}^K \pi_k \mathrm{Tr}(C_k) , \tag{D.7}$$

$$\Delta_\theta^+ := 2\mathrm{e}^{-1} \sum_{k=1}^K \pi_k \max_{1 \leq j \leq 2} \lambda_j^{(k)} . \tag{D.8}$$

Démonstration. By (3.27), if Tr denotes the Trace operator,

$$\begin{aligned}\triangle H_\theta(X) &= \mathrm{Tr}\left[\nabla \alpha_\theta(x) \right] , \\ &= -\mathrm{Tr}\left[\nabla \left(\sum_{k=1}^K \pi_k \varphi_k(\gamma x)\gamma^{-1} C_k(\gamma x - \mu_k) \right) \right] , \\ &= -\mathrm{Tr}\left[\sum_{k=1}^K \left(\pi_k \nabla\varphi_k(\gamma x)\gamma^{-1} \left[C_k(\gamma x - \mu_k) \right]^T + \pi_k \, \varphi_k(\gamma x)\gamma^{-1} C_k \gamma \right) \right] , \\ &= -\sum_{k=1}^K \pi_k \varphi_k(\gamma x) \left\{ -\mathrm{Tr}(\left[C_k(\gamma x - \mu_k) \right] \left[C_k(\gamma x - \mu_k) \right]^T) + \mathrm{Tr}(C_k) \right\} , \\ &= \underbrace{\sum_{k=1}^K \pi_k \varphi_k(\gamma x) \| C_k(\gamma x - \mu_k) \|^2}_{I(x)} - \underbrace{\sum_{k=1}^K \pi_k \varphi_k(\gamma x) \mathrm{Tr}(C_k)}_{J(x)} .\end{aligned}$$

By definition of φ_k, for all $1 \leq k \leq K$,

$$0 \leq J(x) \leq \underbrace{\sum_{k=1}^{K} \pi_k \mathrm{Tr}(C_k)}_{-\triangle_\theta^-}.$$

Following the same steps as for the proof of Lemma D.2.1,

$$0 \leq I(x) \leq \underbrace{2\mathrm{e}^{-1} \sum_{k=1}^{K} \pi_k \max_{1 \leq j \leq 2} \lambda_j^{(k)}}_{\triangle_\theta^+}.$$

□

D.3 Unbiased likelihood estimation for model selection

In the numerical Section, we need to estimate the loglikelihood $L(\mathbf{Y}^g, \theta)$ of each path $1 \leq g \leq G$ for a given parameter θ in order to : a) choose the best estimate among the ones obtained from different starting points in the EM procedure and b) compute the approximate AIC criterion for the chosen estimate to select the best number of components in our model. Note that,

$$L(\mathbf{Y}^g, \theta) = \sum_{j=0}^{n^g} \log p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta),$$

where $p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta)$ is the conditional distribution of Y_j^g (at time t_j^g) given Y_{j-1}^g (at time t_{j-1}^g) when the parameter value is θ (with the convention that $p_{t_0^g - t_{-1}^g}(Y_0^g, Y_{-1}^g, \hat{\theta})$ is the likelihood of the first observation Y_0^g of path g). By (Beskos *et al.*, 2009, Theorem 1), using the EA1 algorithm, it is possible, for any $1 \leq g \leq G$ and any $1 \leq j \leq n^g$, to define an unbiased estimator of $p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta)$. Following the same steps as in (Beskos *et al.*, 2009, Theorem 1), it can be proved that

$$\varphi_{t_j^g - t_{j-1}^g}(Y_j^g - Y_{j-1}^g) \exp\left\{H_\theta(Y_j^g) - H_\theta(Y_{j-1}^g)\right\} \times \prod_{i=1}^{\Upsilon_\theta} (1 - \phi_\theta(\omega_{U_i})/\Lambda_\theta) \tag{D.9}$$

is an unbiased estimator of $p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta)$ when ω has the law of a Brownian bridge between (t_{j-1}^g, Y_{j-1}^g) and (t_j^g, Y_j^g), Υ_θ is a Poisson random variable with mean $\Lambda_\theta(t_j^g - t_{j-1}^g)$ and the $(U_i)_{1 \leq i \leq \Upsilon}$ are i.i.d. uniform random variables on $[0, t_j^g - t_{j-1}^g]$.

Proof of equation (D.9)

To simplify notations, assume that $t_{j-1} = 0$ and t_j=T. Let $\mathbb{Q}_\theta^{T,x,y}$ (resp. $\mathbb{W}^{T,x,y}$) be the probability measure induced by $(Y_s)_{0 \leq s \leq T}$, solution to equation (3.28), (resp. $(W_s)_{0 \leq s \leq T}$) on the Wiener space, conditioned on hitting x at $t = 0$ and y at $t = T$. By Girsanov theorem, for

ω an element of C,

$$\frac{d\mathbb{Q}_\theta^{T,x,y}}{d\mathbb{W}^{T,x,y}}(\omega) = \frac{\varphi_T(y-x)}{p_{\theta,T}(x,y)} \exp\left(H_\theta(y) - H_\theta(x) - \int_0^T \phi_\theta(\omega_s)ds\right) \tag{D.10}$$

Where H_θ is defined in (3.26), ϕ_θ in (D.2), φ_T is the p.d.f of a zero mean bivariate Gaussian distribution with covariance $T \times I_2$ and $p_{\theta,T}(x,y)$ is the p.d.f. of the distribution of Y_T conditioned on $Y_0 = x$. By computing the expectation of each term of equation (D.10) with respect to $\mathbb{W}^{T,x,y}$, we get

$$p_{\theta,T}(x,y) = \varphi_T(y-x) \exp\left(H_\theta(y) - H_\theta(x)\right) \times \mathbb{E}_{\mathbb{W}^{T,x,y}}\left[\exp\left(-\int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right)\right] \tag{D.11}$$

Let Λ_θ be a positive number such as in (D.4) and $\omega := (\omega_s)_{0\leq s\leq T}$ be such that $\omega \sim \mathbb{W}^{T,x,y}$. Then, the exponential term in (D.11) may be written

$$\exp\left(-\int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right) = \mathrm{e}^{-\Lambda_\theta T} \exp\left(\Lambda_\theta T - \int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right), \tag{D.12}$$

$$= \mathrm{e}^{-\Lambda_\theta T} \sum_{k\geq 0} \frac{\left(\Lambda_\theta T - \int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right)^k}{k!}. \tag{D.13}$$

Let U be a random variable with a uniform distribution on $[0,T]$, independent of ω, then the following equality holds

$$\frac{1}{T}\int_0^T \phi_\theta(\omega_s)\mathrm{d}s = \mathbb{E}\left[\phi_\theta(\omega_U)|\omega\right]. \tag{D.14}$$

Then, if the random variables U_i are independent and with the same uniform distribution on $[0,T]$,

$$\exp\left(-\int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right) = \mathrm{e}^{-\Lambda_\theta T} \sum_{k\geq 0} \frac{\prod_{i=1}^k \mathbb{E}\left[\Lambda_\theta T - T\phi_\theta(\omega_{U_i})|\omega\right]}{k!}. \tag{D.15}$$

By independence of the random variables U_i,

$$\exp\left(-\int_0^T \phi_\theta(\omega_s)\mathrm{d}s\right) = \mathrm{e}^{-\Lambda_\theta T}\mathbb{E}\left[\sum_{k\geq 0} \frac{\prod_{i=1}^k \left(\Lambda_\theta T - T\phi_\theta(\omega_{U_i})\right)}{k!}\middle|\omega\right], \tag{D.16}$$

$$= \mathbb{E}\left[\sum_{k\geq 0} \mathrm{e}^{-\Lambda_\theta T}\frac{(\Lambda_\theta T)^k}{k!}\prod_{i=1}^k\left(1 - \frac{\phi_\theta(\omega_{U_i})}{\Lambda_\theta}\right)\middle|\omega\right]. \tag{D.17}$$

The proof is concluded upon noting that if Υ_θ is a Poisson random variable with parameter $\Lambda_\theta T$, independent of ω and U_i for all i,

$$\exp(-\int_0^T \phi_\theta(\omega_s)ds) = \mathbb{E}\left[\prod_{i=1}^{\Upsilon_\theta}(1 - \frac{\phi_\theta(\omega_{U_i})}{\Lambda_\theta})\middle|\omega\right]. \tag{D.18}$$

Expression of the unbiased estimator

Then, an unbiased Monte Carlo estimate of $p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta)$ may be obtained by N independent realizations of these random variables. However, as the Poisson mean depends on θ, we would have to sample all the random variables for each value of the parameter θ. Nevertheless, in our case we need to estimate the likelihood only for a finite number of parameter values obtained at the end of the EM algorithm for each starting point : $\theta \in \{\hat{\theta}_i\}_{1 \leq i \leq p}$. In this case, writing $\bar{\Lambda} := \max_{1 \leq i \leq p} \Lambda_{\hat{\theta}_i}$, the unbiased estimator of $p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta)$ we may use is given by

$$\varphi_{t_j^g - t_{j-1}^g}(Y_j^g - Y_{j-1}^g) \exp\left\{H_\theta(Y_j^g) - H_\theta(Y_{j-1}^g)\right\} \times \prod_{i=1}^{\Upsilon} \left(1 - \phi_\theta(\omega_{U_i})/\bar{\Lambda}\right) ,$$

where ω has the law of a Brownian bridge between (t_{j-1}^g, Y_{j-1}^g) and (t_j^g, Y_j^g), Υ is a Poisson random variable with mean $\bar{\Lambda}(t_j^g - t_{j-1}^g)$ and the $(U_i)_{1 \leq i \leq \Upsilon}$ are i.i.d. uniform random variables on $[0, t_j^g - t_{j-1}^g]$. As $\bar{\Lambda}$ is independent of θ, drawing once

- $(\omega^k)_{1 \leq k \leq N}$, N independent Brownian bridges between (t_{j-1}^g, Y_{j-1}^g) and (t_j^g, Y_j^g) ;
- $(\Upsilon_k)_{1 \leq k \leq N}$, N independent Poisson random variables with mean $\bar{\Lambda}(t_j^g - t_{j-1}^g)$;
- $\{(U_i^k)_{1 \leq i \leq \Upsilon}\}_{1 \leq k \leq N}$ independent uniform random variables on $[0, t_j^g - t_{j-1}^g]$;

allows to define the following estimator of $p_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta)$ for all values $\theta \in \{\hat{\theta}_i\}_{1 \leq i \leq p}$:

$$\begin{aligned} p^N_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta) :=& \varphi_{t_j^g - t_{j-1}^g}(Y_j^g - Y_{j-1}^g) \exp\left\{H_\theta(Y_j^g) - H_\theta(Y_{j-1}^g)\right\} \\ & \times \frac{1}{N} \sum_{k=1}^{N} \prod_{i=1}^{\Upsilon_k} \left(1 - \phi_\theta(\omega^k_{U_i^k})/\bar{\Lambda}\right) . \end{aligned} \tag{D.19}$$

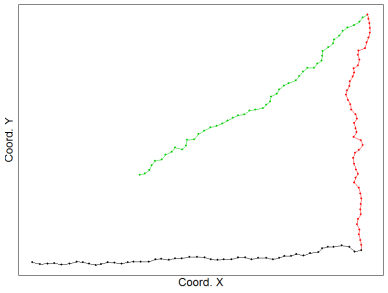
Finally, choosing N, the estimator of the likelihood is then computed with the equation

$$L^N(\mathbf{Y}^g, \theta) = \sum_{j=0}^{n^g} \log p^N_{t_j^g - t_{j-1}^g}(Y_j^g, Y_{j-1}^g, \theta) , \tag{D.20}$$

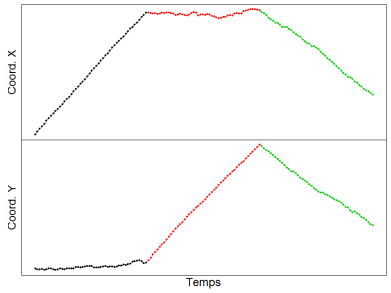
Annexe E

Figures complémentaires

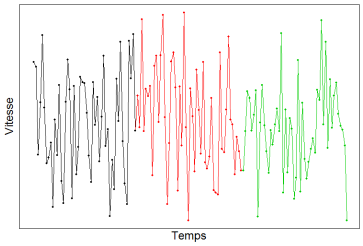
E.1 Figures complémentaires à la section 2.1.1



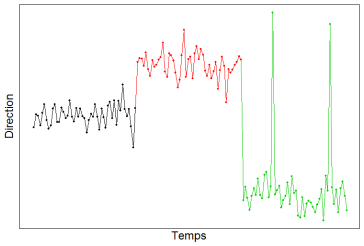
(a) Trajectoire



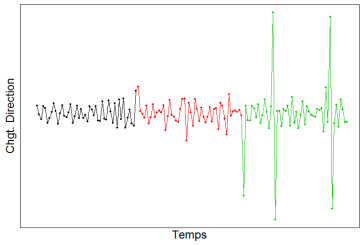
(b) Positions



(c) Vitesses

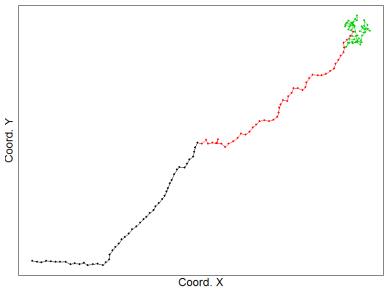


(d) Directions

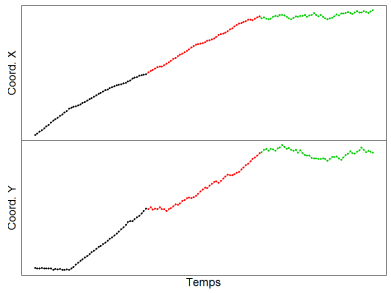


(e) Changements de direction

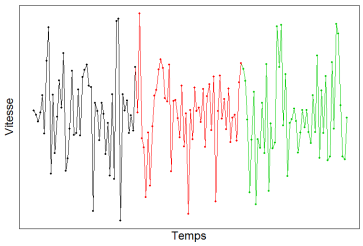
FIGURE *E.1 – Exemple de trajectoire simulée, mélange de 3 comportements. Chaque comportement est caractérisé par une destination différente. Ici, c'est la série des directions qui semble la plus adéquate pour caractériser le comportement*



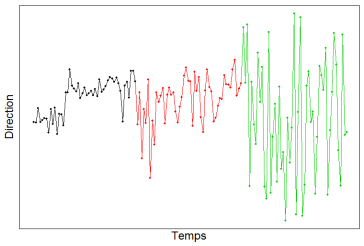
(a) Trajectoire



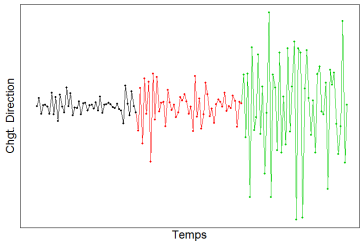
(b) Positions



(c) Vitesses



(d) Directions



(e) Changements de direction

FIGURE *E.2 – Exemple de trajectoire simulée, mélange de 3 comportements. Chaque comportement est caractérisé par une diffusion croissante. Ici, c'est la série des changements de direction qui semble la plus adéquate pour caractériser le comportement*

E.2 Figures complémentaires à la section 3.3, application du modèle GaP

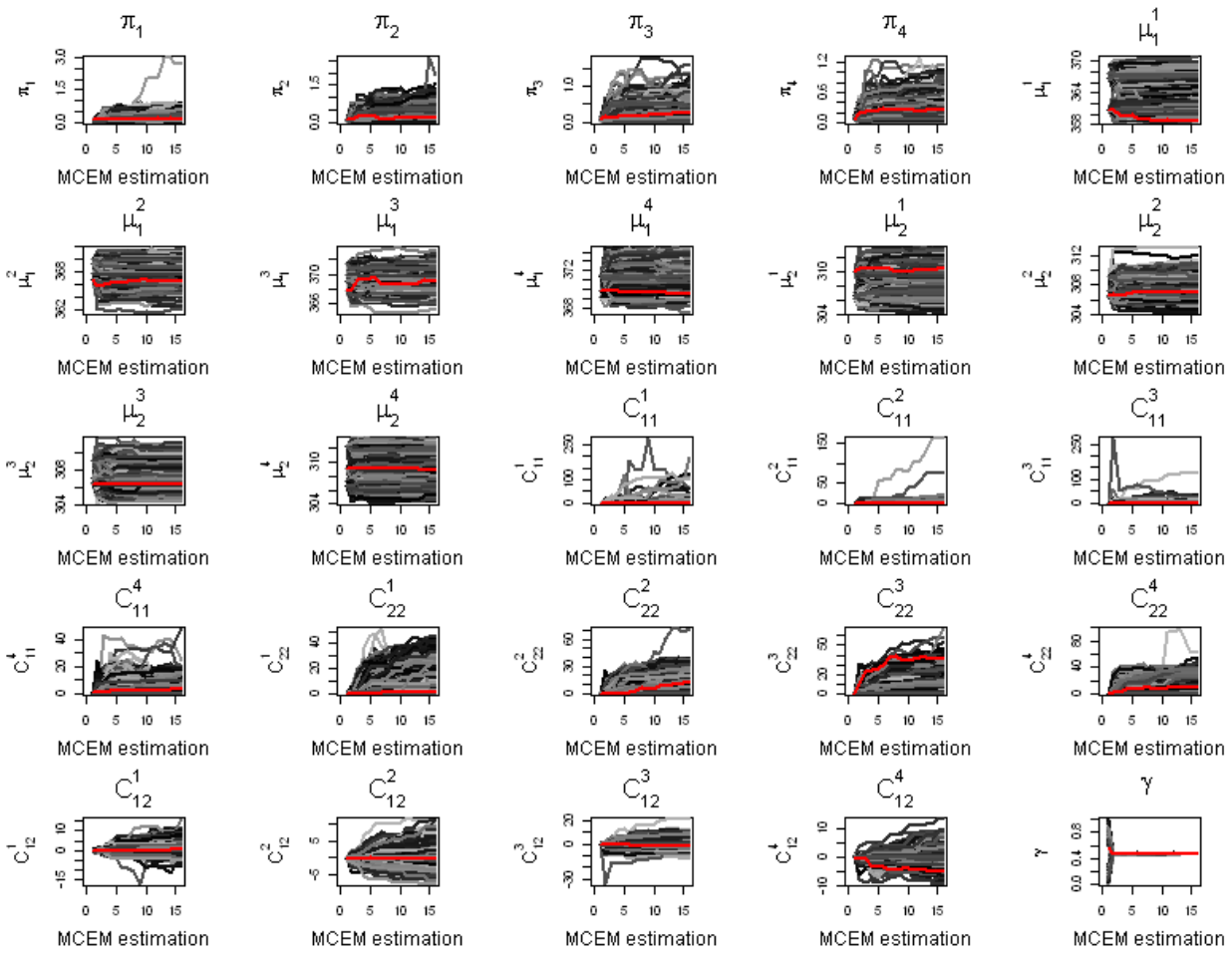


FIGURE *E.3 – Trajectoires de l'algorithme MCEM pour l'estimation du modèle GaP en zone de Boulogne. Les trajectoires sont ici estimées pour $K = 4$ La trajectoire rouge est celle ayant le meilleur point final au sens de l'approximation de la vraisemblance. Pour la plupart des trajectoires, la convergence advient rapidement (moins de 10 itérations). Il semble donc plus pertinent de privilégier le nombre de points de départ au nombre d'itérations.*