Habilitation à diriger des recherches

université PARIS-SACLAY

Etude probabiliste de la biodiversité et de la génétique des populations à reproduction sexuée

Habilitation à diriger des recherches de l'Université Paris-Saclay

présentée et soutenue à Palaiseau, le 04/02/2026, par

Camille CORON

Composition du jury

Vincent BANSAYE Professeur, Ecole Polytechnique	Rapporteur
Alison ETHERIDGE Professeur, Oxford University	Rapportrice
Peter PFAFFELHUBER Professeur, Albert-Ludwigs Universität, Freiburg	Rapporteur
Pierre BARBILLON Professeur, AgroParisTech	Examinateur
Arnaud ESTOUP Directeur de recherche, INRAE	Examinateur
Emmanuelle PORCHER Professeur, Museum National d'Histoire Naturelle	Examinatrice
Stéphane ROBIN Professeur, Sorbonne Université	Examinateur

Table des matières

Remerciements

Je vous promets qu'aucune des deux prochaines pages n'a exploité de grand modèle de langage – autre que le mien, qui n'est malheureusement pas si grand que ça.

Mes premiers remerciements vont spontanément à mes co-auteurs. Avec eux et grâce à eux j'ai appris ou réappris, et pratiqué, l'étude des processus de branchement, la convergence des suites de processus stochastiques vers des solutions de systèmes dynamiques, la caractérisation des mesures stationnaires de chaînes de Markov à l'aide d'arbres couvrants, les modèles linéaires généralisés, la confrontation des modèles aux données "réelles" (il paraît que toutes les données le sont...), les bases de Python, de R, de BUGS, et le fonctionnement des systèmes de reproduction des plantes. J'ai même réappris (à plusieurs reprises !) l'incontournable théorème de Perron-Frobenius. Avec eux (mais pas à cause d'eux) j'ai été parfois énervée, frustrée, déçue, mais toujours avec eux j'ai surtout beaucoup calculé, beaucoup réfléchi, beaucoup ri, et partagé l'étonnant bonheur qu'est la découverte d'un résultat et d'une preuve mathématiques. J'ai notamment une reconnaissance particulière pour Christophe Giraud qui m'a initiée au monde des statistiques et des données biologiques. J'ai eu une chance inouïe de pouvoir faire ce saut après ma thèse, en étant accompagnée et qui plus est en ayant accès à des données qui avaient le bon goût d'illustrer nos résultats théoriques (ce qui n'est en général pas leur première qualité comme j'ai pu avec amertume le découvrir par la suite). Je remercie aussi Yves Le Jan qui a travaillé avec moi sur des sujets qui m'intéressaient, et a supporté avec patience mon manque de disponibilité. Charline, Hélène, Manon [1], je vous remercie énormément pour nos journées de travail, à Paris, à Lyon, à Grenoble, à Clermont-Ferrand, pour nos délicieux restaurants et pour nos voyages détente et retrouvailles, à Avignon (cf photo ci-contre), à Barcelone. Je garde un souvenir particulièrement ému de notre séjour à Grenoble en plein Covid, entre couvre-feu et semi-confinement, dans un Airbnb réconfortant, et de "Drunk" dans un cinéma avignonnais qui venait de rouvrir. À quand Amsterdam ?

C'est grâce à Sylvie Méléard que j'ai rencontré ces trois chercheuses inspirantes. Je l'en remercie, tout comme je la remercie de m'avoir, par son encadrement durant ma thèse, parfaitement formée et lancée dans ce domaine des mathématiques appliquées, notamment à la génétique.

Je suis infiniment reconnaissante à Vincent Bansaye, Alison Etheridge et Peter Pfaffelhuber d'avoir passé du temps à lire et évaluer mon manuscrit. Je sais que leur emploi du temps est chargé et je suis à la fois gênée d'avoir sollicité leur aide et honorée qu'ils aient accepté ma demande. Leurs commentaires et questions sont très intéressants et utiles pour moi. Je remercie aussi chaleureusement Emmanuelle Porcher, Stéphane Robin, Pierre Barbillon et Arnaud Estoup d'avoir accepté de participer à mon jury de soutenance.

Je remercie aussi Arnaud Becheler, Emma Thulliez, Luce Breuil, Lucas Machado Moschen,

1. Charline Smadi, Hélène Leman, Manon Costa

Gaspard Dousson-Lys, Léo Micollet, Violette Kubiaczyk et Lucas Rey qui m'ont fait confiance pour encadrer leur recherche.

J'ai été pendant 9 ans maîtresse de conférence au Laboratoire de Mathématiques d'Orsay et à l'IUT de Sceaux. J'ai reçu un accueil très chaleureux dans ces deux lieux, j'y ai été encouragée, je m'y suis sentie bien. Je remercie notamment Sara Brofferio et Virginie Demulier pour leur générosité et leur intelligence relationnelle. Elles m'ont toujours très bien conseillée et ont représenté de véritables modèles d'enseignantes-chercheuses pour moi. À Orsay j'ai vécu un changement de poste, des arrivées, un changement de bâtiment, d'autres arrivées, des départs, puis mon départ. Je garde de ce lieu des souvenirs amusés de repas passés entre Elisabeth Gassiat et Jean-François Le Gall qui parlaient théâtre et philharmonie, des souvenirs joyeux de pots de thèse, repas de conférences, et autres galettes des rois, des souvenirs heureux d'un bureau partagé avec Jean-Michel Poggi, des souvenirs passionnants de collaborations intéressantes et variées. Merci à tous, et notamment à Vincent Rivoirard pour sa gentillesse, ses conseils et ses explications toujours intéressantes sur le monde de la recherche ! Quitter ce laboratoire et les relations que j'y avais tissées a été très difficile. Fort heureusement en montant tout simplement la colline qui se trouvait juste à côté j'ai atterri à MIA Paris-Saclay, dans une équipe, un institut, une école qui m'ont accueillie merveilleusement bien, et semblent m'accepter malgré mes sujets qui sont devenus d'un seul coup terriblement théoriques. Je me sens à ma place ici et j'espère bien pouvoir rester ! ;-) Merci déjà pour l'accueil et pour ces premières années très joyeuses !

A divers titres je peux dire que sans mes parents je n'écrirais pas ces lignes. Je les remercie infiniment pour l'amour et l'éducation qu'ils m'ont donnés. Je remercie aussi mes frère et soeurs, notamment la vice-présidente de l'Université Paris-Saclay, Clotilde Coron, qui me fait l'amitié de suivre avec attention mes progrès depuis que j'ai 3 ans. Je remercie Jennifer et Aliénor, qui me permettent de ne pas me mettre à bafouiller quand on me demande si les enfants de mes parents sont tous chercheurs. Je remercie enfin Basile de suivre mes pas, déjà bien tracés avant moi, en ayant le courage de se lancer dans une carrière de chercheur en mathématiques. Être proche d'eux et de leurs familles me rend très heureuse.

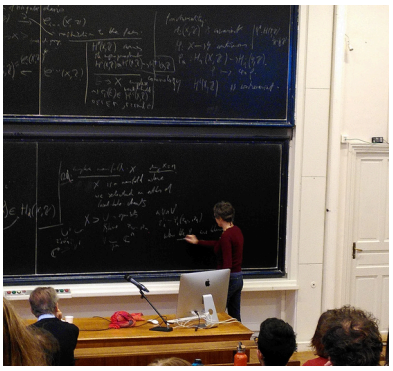
Je remercie aussi Diala, Émilie, Agathe, qui me sortent (trop rarement !) de mon quotidien, m'emmènent à l'escalade, au théâtre, partagent ma folie pour les activités manuelles, et des petits morceaux de leurs vies avec moi.

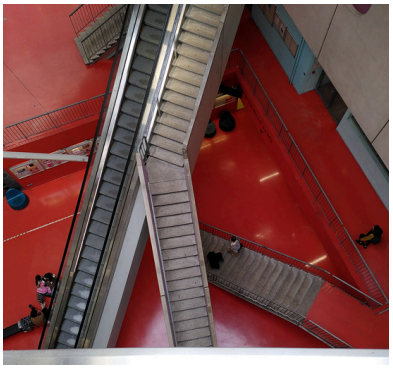
Pour finir, je remercie Clément pour son amour et son soutien constants, Marie, Antoine et Lucie pour leurs rires, leur énergie débordante, et pour leur simple existence. Je leur souhaite d'avoir un métier aussi passionnant que le mien.

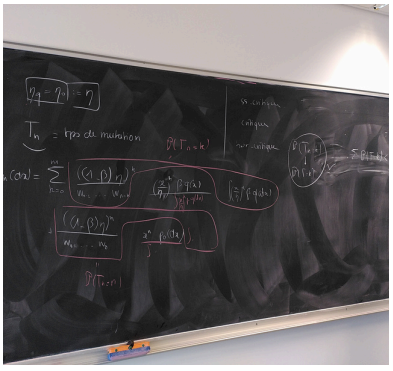
J'aime prendre en photo certains instants : de beaux paysages, des situations incongrues, des moments partagés. Je vous laisse donc avec quelques photos non exhaustives de lieux que j'ai aimé découvrir durant ces dernières années de travail.

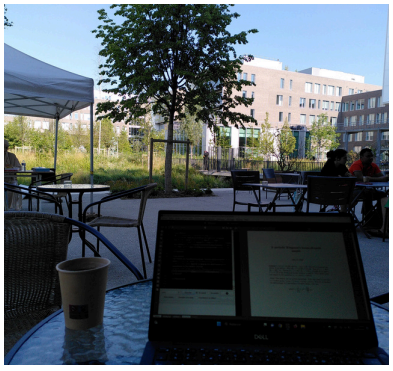
Camille


MATHEMATIQUE









Publications

1. E. Thulliez, C. Coron (2025) Local improvement of NO2 concentration maps derived from physicochemical models, using low-cost sensors, ArXiv 2505.17564.
2. C. Coron, Y. Le Jan (2025) Genetic contribution of advantaged ancestors in the biparental Moran model – finite selection. ArXiv 2502.01178.
3. B. Auder, C. Coron, J.-.M. Poggi, E. Thulliez (2025) Debiasing physico-chemical models in air quality monitoring by combining different pollutant concentration measurements. Communications in Statistics : Case Studies, Data Analysis and Applications, 1–26.
4. C. Coron, Y. Le Jan (2024) Genetic contribution of an advantaged mutant in the biparental Moran model - finite selection. ArXiv 2405.08404.
5. C. Coron, Y. Le Jan (2024) Genetic Contribution of an Advantaged Mutant in the Biparental Moran Model. Ukrainian Mathematical Journal 75, 1666–1672.
6. C. Coron, M. Costa, H. Leman, V. Llaurens, C. Smadi (2022) Origin and persistence of polymorphism in loci targeted by disassortative preference: a general model. Journal of Mathematical Biology 86(1) :4.
7. C. Coron, Y. Le Jan (2022) Pedigree of the biparental Moran model Journal of Mathematical Biology 84, 51.
8. C. Coron, M. Costa, F. Laroche, H. Leman, C. Smadi (2021). Emergence of homogamy in a two-loci stochastic population model. ALEA 18, 469–508.
9. A. Becheler, C. Coron, S. Dupas (2019) : The Quetzal Coalescence template library: A C++ programmers resource for integrating distributional, demographic and coalescent models. Molecular Ecology Resources 19(3) :788–793.
10. D. Abu Awad, C. Coron (2018) : Effects of demographic stochasticity and life-history strategies on times and probabilities to fixation: an individual-based model. Heredity 121(4) :374–386.
11. C. Coron, S. Méléard, D. Villemonais (2018) : Impact of demography on extinction/fixation events Journal of Mathematical Biology 78(3) :549–577.
12. C. Calenge, C. Coron, C. Giraud, R. Julliard (2018) : Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. Environmental and Ecological Statistics 25, 71-–93.
13. C. Coron, M. Costa, H. Leman, C. Smadi (2018) : A stochastic model for speciation by mating preferences. Journal of Mathematical Biology 76, 1421–1463.
14. C. Giraud, C. Calenge, C. Coron et R. Julliard (2015) : Capitalizing on opportunistic data for monitoring species relative abundances. Biometrics 72 : 649–658.
15. C. Coron (2015) : A model for Mendelian populations demogenetics. ESAIM : proceedings 51 :122–132.

16. C. Coron (2015) : Slow-fast stochastic diffusion dynamics and quasi-stationary distributions for diploid populations. Journal of Mathematical Biology 72(1-2) :171–202.
17. C. Coron (2014) : Stochastic modeling of density-dependent diploid populations and extinction vortex. Advances in Applied Probability 46, 446–477.
18. C. Coron, S. Méléard, E. Porcher et A. Robert (2013) : Quantifying the mutational meltdown in diploid populations. American Naturalist 181(5) : 623–636.

Ce manuscrit est pour l'essentiel une synthèse des articles 2, 4, 5, 6, 7, 8, 12, 13, 14.

Introduction

Ma recherche se situe en probabilités pour la biologie. Je suis particulièrement intéressée par la modélisation et l'analyse probabiliste de la diversité génétique, de la biodiversité, et de leurs dynamiques temporelles et spatiales. Ce manuscrit, après une introduction destinée à expliquer mes motivations, mes approches et le contexte général de mes recherches, est divisé en trois parties qui portent sur des questions biologiques différentes, abordées en développant des modèles mathématiques différents, et avec des collaborateurs différents.

Approche générale, motivation et parcours Mon travail consiste à répondre à une question biologique à l'aide de modèles mathématiques que je crée puis que j'étudie. Une des difficultés de ce domaine est de définir une bonne question biologique, c'est-à-dire une question qui intéresse les biologistes, qui soit bien posée, et à laquelle on puisse répondre à l'aide de modèles suffisamment simples pour être étudiés mathématiquement. Bien que le vivant présente en général des comportements complexes, c'est justement cette simplicité qui doit permettre de capter l'essence des phénomènes biologiques considérés, mais aussi d'obtenir des résultats et de développer des techniques mathématiques qui puissent présenter un intérêt aussi pour la recherche en mathématiques. Enfin, la prise en compte des données, leur modélisation et leur utilisation pour mieux modéliser et comprendre le vivant est aussi un enjeu très important de mon domaine de recherche, que j'essaye d'aborder. J'ai travaillé souvent avec des biologistes, toujours avec des mathématiciens, et nous avons essayé de réaliser des travaux qui puissent concerner ces deux communautés.

Durant ma thèse au Centre de Mathématiques Appliquées de l'École Polytechnique, sous la direction de Sylvie Méléard, je me suis intéressée à la modélisation et l'étude probabiliste de l'évolution génétique des populations à reproduction sexuée. J'ai ensuite occupé pendant un an un poste de Lectrice Hadamard, au Laboratoire de Mathématiques d'Orsay. Durant cette année j'ai eu la chance immense de découvrir un domaine et une communauté complètement différents, en travaillant avec Christophe Giraud notamment sur la combinaison de jeux de données scientifiques et citoyens afin de mieux évaluer l'état et la dynamique de la biodiversité. Les travaux que j'ai menés par la suite, au Laboratoire de Mathématiques d'Orsay en tant que maîtresse de conférence, puis dans l'unité Mathématiques et Informatique Appliquées de Paris-Saclay (INRAE) à AgroParisTech en tant que professeur junior, ont pour la plupart un lien fort avec l'un de ces deux domaines de probabilité appliquées à la biologie : l'évolution génétique des popula-

tions à reproduction d'une part, et la combinaison de données environnementales, ou biologiques, d'autre part. Je vais maintenant en présenter une partie, ainsi que les perspectives de recherche et d'encadrement que j'envisage pour les prochaines années.

Impacts de la reproduction sexuée et de la démographie sur la diversité génétique La composition génétique d'une population à reproduction sexuée (c'est-à-dire l'ensemble des génomes de tous les individus qui la composent) est le résultat de processus très complexes et qui ne sont pas indépendants les uns des autres. Parmi eux on notera par exemple le choix d'un partenaire de reproduction et donc la construction progressive d'une population munie d'un graphe de parenté, la transmission du génome le long de ce graphe et les mutations ayant lieu lors de cette transmission, la traduction de ce génome en termes de capacité de survie des individus, et de reproduction des couples d'individus. Étudier la composition génétique de ces populations est un problème difficile et important, qui peut-être abordé de façons multiples, allant d'approches très théoriques à des techniques beaucoup plus appliquées et proches des données.

Durant ma thèse je me suis intéressée à la dynamique de la composition génétique d'une population à reproduction sexuée et de taille variable, dans laquelle les individus sont caractérisés par leur génome à un seul locus diploïde et bi-allélique (Coron (2015)). J'ai pour cela créé et étudié des modèles probabilistes individus-centrés, plus précisément des processus de naissance et mort avec interactions, qui sont caractérisés par un certain nombre de paramètres démographiques, qui déterminent la reproduction et la mort des individus. J'ai alors montré l'existence d'échelles lentes-rapides dans la dynamique de ces processus, et étudié leur comportement quasi-stationnaire.

Par la suite j'ai continué à me passionner pour ce sujet qui est très riche, mais avec des approches, des questions, et des collaborateurs différents. Cette partie de mon travail peut se découper en 3 thématiques : étude de l'interaction entre démographie et diversité génétique, étude de la composition génétique d'une population biparentale, et étude du rôle de la reproduction sexuée dans l'évolution génétique des populations. Je me suis penchée sur l'interaction réciproque entre démographie et diversité génétique dans le cas de populations à reproduction sexuée, d'une part avec Sylvie Méléard et Denis Villemonais (Coron *et al.* (2019)), et d'autre part avec Diala Abu Awad (Abu Awad et Coron (2018)). Dans Coron *et al.* (2019), nous avons montré que le comportement quasi-stationnaire de la diversité génétique d'une population est caractérisé par l'intégrabilité d'un processus de diffusion stochastique de dimension 1, qui est elle-même assurée par un critère explicite sur les paramètres démographiques. Dans Abu Awad et Coron (2018) nous avons étudié l'impact de ces paramètres démographiques (liés à la notion de traits d'histoires de vie, définie par les généticiens des populations pour étudier l'évolution darwinienne des populations, Flatt et Heyland (2011)) sur la vitesse de fixation d'allèles et donc sur la perte de diversité génétique. Ensuite, j'ai étudié, avec Yves Le Jan (Coron et Le Jan (2022, 2024a,b)), la composition génétique d'une population à reproduction sexuée. Dans cette série de travaux, présentée dans le Chapitre 1 de ce manuscrit, nous étudions la proportion asymptotique du génome d'une population biparentale, qui provient d'un ancêtre donné. Notre travail s'inscrit ainsi dans

la lignée de l'article Derrida *et al.* (2000) et s'ajoute aux travaux Chang (1999); Lambert *et al.* (2018); Newman *et al.* (2024) qui explorent la structure génétique des populations biparentales avec des approches très différentes, intéressantes et complémentaires. Nous prouvons en particulier que dans un modèle de Moran biparental neutre, la contribution asymptotique d'un ancêtre au génome de la population considérée est soit égale à 0 (avec probabilité 1/2, ce qui signifie que l'ancêtre a une descendance non éternelle), soit suit une loi exponentielle de paramètre 1/2 (Théorème 1.1). En ajoutant de la sélection à ce modèle nous analysons l'impact de la sélection sur la proportion de génome transmise par un individu. A titre d'exemple, nous montrons que si une population est initialement constituée de 1% d'individus très favorisés génétiquement, alors ils seront en temps long en moyenne à l'origine de 19% du génome de la population (Théorèmes 1.2 et 1.3). Enfin, avec Manon Costa, Hélène Leman et Charline Smadi (Coron *et al.* (2018b, 2021, 2022)), j'ai étudié l'impact des préférences d'appariement sur la spéciation et la diversité génétique. Dans ces travaux, menés en partie en collaboration avec les biologistes de l'évolution Fabien Laroche et Violaine Llaurens, et présentés dans le Chapitre 2 de ce manuscrit, nous nous intéressons plus précisément à l'homogamie (le fait pour un individu de se reproduire plus facilement avec un individu qui lui ressemble) et à l'hétérogamie. Nous déterminons notamment les conditions d'émergence de l'homogamie (Théorème 2.3), comment l'homogamie peut engendrer la spéciation (Théorème 2.7) et la quantité de diversité génétique permise par l'hétérogamie (Théorèmes 2.8 et 2.9). Nos travaux constituent une nouvelle façon d'étudier mathématiquement le rôle des préférences d'appariement dans l'évolution génétique des populations et la spéciation. En particulier le fait de considérer des modèles stochastiques individu-centrés permet d'étudier des quantités importantes, comme la probabilité d'invasion d'un mutant homogame dans une population, déterminée dans le Théorème 2.3.

Amélioration du suivi de la biodiversité par combinaison de données citoyennes et scientifiques Le développement et l'étude de modèles probabilistes visant à une meilleure compréhension des systèmes biologiques me passionne. Toutefois, pour que ce travail me paraisse véritablement pertinent, il est essentiel pour moi d'avoir une connaissance des données et de concevoir des modèles permettant de les exploiter pour répondre à des questions biologiques. Durant mon année de post-doctorat au laboratoire de Mathématiques d'Orsay j'ai travaillé avec Christophe Giraud, Romain Julliard, et Clément Calenge, sur la question de l'exploitation de données issues de programmes de sciences citoyennes, afin d'améliorer l'évaluation et le suivi de la biodiversité. Les programmes de sciences citoyennes, ou sciences participatives, sont mis en place par des scientifiques, mais utilisent le temps, l'énergie et la bonne volonté des citoyens afin de récolter des observations. Ils sont en général caractérisés par un protocole d'observation assez léger, voire inexistant, mais aussi par un très grand nombre de données récoltées, ce qui incite à chercher une façon de les calibrer pour les utiliser de façon pertinente. Dans le cas qui nous a intéressés nous avions à notre disposition deux jeux de données d'observations d'oiseaux en Aquitaine : l'un récolté par des professionnels selon un protocole précis (avec notamment une information du temps passé à observer et la consigne pour les observateurs de rapporter

toutes leurs observations) et l'un issu d'observations faites par des citoyens, sans contrainte particulière. Notre but était, à partir de ces jeux de données, de fournir des cartes d'abondances relatives d'espèces, c'est-à-dire d'être capable de comparer le nombre d'individus d'une espèce donnée, à deux endroits différents de l'espace considéré. Estimer ces abondances relatives est possible à l'aide du seul jeu de données professionnel, du fait du protocole très strict qu'il impose. Néanmoins les estimations auxquelles il conduit sont très bruitées, car il contient peu de données. Notre approche a consisté à coupler, au travers d'un modèle probabiliste, ces deux jeux de données, de façon à bénéficier à la fois de la calibration des données scientifiques et de l'abondance des données citoyennes, et améliorer ainsi la précision des estimations d'abondances relatives d'espèces obtenues avec le seul jeu de données professionnel (Théorème 3.1). Notre approche est originale : la plupart des travaux visant à calibrer les données issues de programmes de sciences participatives consistent plutôt à essayer de les débiaiser en estimant le temps passé par les observateurs sur le terrain ou plus grossièrement en remplaçant ce temps par le nombre d'observateurs, ou le nombre d'observations. Ce travail ouvre de nombreuses perspectives, car cette situation dans laquelle plusieurs jeux de données sont issus d'une même réalité biologique est de plus en plus fréquente.

Perspectives Mes recherches actuelles et perspectives de recherche sont développées à la fin de chaque chapitre. Elles sont centrées sur l'étude de l'évolution génétique et de la dynamique de populations à reproduction sexuée, mais portent sur des questions et applications biologiques différentes, abordées avec des approches mathématiques aussi très variées. J'aimerais notamment comprendre l'impact de l'hybridation et de la structure des familles sur le génome de populations à reproduction sexuée. Je cherche par ailleurs à modéliser la dynamique de populations d'insectes contrôlées par la technique de l'insecte stérile, et à étudier l'optimisation de cette technique. Je voudrais enfin estimer la démographie et l'histoire migratoire de populations à partir de données génétiques, de données de comptages, ou par combinaison de ces différents types de données. Les applications de ces différentes questions se situent en agronomie, génétique animale, histoire de l'Homme et santé.

Chapitre 1

Génétique des populations biparentales

1.1 Introduction

Cette partie de mon travail a été réalisée en collaboration avec Yves Le Jan (Université Paris-Saclay). Notre objectif était d'étudier la composition génétique des populations à reproduction sexuée, donc les populations dans lesquelles le génome d'un individu est une fonction aléatoire du génome de ses deux parents. Cette question de recherche est très importante, par exemple pour comprendre l'avantage conféré aux espèces par la reproduction sexuée (Agrawal (2001)), mais surtout pour inférer, à partir de données génomiques, l'histoire démographique ainsi que certains paramètres régissant la dynamique des populations et l'histoire de vie des individus, comme les paramètres de sélection ou les préférences d'appariement. Cet objectif d'inférence est notamment à l'origine du développement et de l'étude des coalescents séquentiellement Markoviens qui sont des modèles approchés et très étudiés, de graphes ancestraux avec recombinaison (McVean et Cardin (2005)). Malgré ces enjeux, la génétique des populations à reproduction sexuée a fait l'objet de relativement peu d'articles portant sur l'étude de modèles probabilistes exacts, et reste un domaine à découvrir. On peut néanmoins en distinguer quelques-uns : les deux articles Chang (1999) et Linder (2009) ont une approche "backward in time", au travers de laquelle ils étudient notamment deux temps en remontant dans le passé d'une population (représentée par un modèle de Wright-Fisher biparental dans le premier article et par un modèle de Moran biparental dans le second) : le premier temps au bout duquel il existe dans la population un ancêtre commun à tous les individus de la population présente, et le premier temps au bout duquel tous les ancêtres sont soit ancêtre de tous les individus de la population présente, soit ancêtre d'aucun individu de la population présente. Étudier les échelles de temps en génétique des populations est essentiel pour comprendre la diversité génétique, qui est assurée par l'accumulation de mutations au cours du temps. L'article Derrida *et al.* (2000) a une approche très différente : les auteurs donnent, sous une hypothèse de grande taille de population, la loi de la proportion asymptotique de génome transmise par un ancêtre donné dans une population modélisée par un modèle de Wright-Fisher biparental. Dans Lambert *et al.* (2018), les auteurs modélisent le génome d'un individu par le

segment $[0,1]$, et ce génome est transmis, avec recombinaison, par chaque couple de parents à leur enfant : en cas de recombinaison le génome d'un des deux parents est coupé à une position uniforme du segment $[0,1]$ et le génome de l'enfant est alors constitué du début du génome d'un de ses parents, suivi de la fin du génome de l'autre. Pour chaque position $x \in [0,1]$, la généalogie d'un échantillon d'individus est alors coalescente, et finit dans un ancêtre commun à tous les individus. Les auteurs s'intéressent alors au coloriage de ce segment en fonction de l'identité de l'ancêtre qui a transmis son génome à l'ensemble de la population, à chaque point x du segment $[0,1]$. Ils donnent la dynamique et la loi stationnaire de ce coloriage. L'article Pfaffelhuber et Wakolbinger (2023) présente une approche plus proche de la génétique quantitative : les auteurs s'intéressent à la transmission des éléments transposables au travers de la reproduction sexuée. Chaque individu porte un certain nombre d'éléments génétiques transposables, et un enfant hérite d'une fonction aléatoire du nombre d'éléments total portés par ses parents. Les auteurs montrent l'existence d'une échelle lente rapide dans ce modèle et caractérisent la distribution stationnaire de la fréquence d'individus portant un certain nombre d'éléments transposables. Enfin, l'article Newman *et al.* (2024) étudie l'impact de l'autofécondation dans le génome des populations à reproduction sexuée et démontre l'importance de prendre en compte le pédigrée dans cette analyse.

Notre travail s'inscrit dans la lignée de l'article Derrida *et al.* (2000), dans le sens où notre approche consiste à étudier la contribution asymptotique d'un ancêtre au génome d'une population.

1.2 Modèle

Base du modèle Nous nous plaçons dans le cadre où la population considérée suit un modèle de Moran biparental avec sélection. Plus précisément, notons N le nombre d'individus dans la population, qui sera constant et sera un paramètre clé du modèle (le seul, dans le cas le plus simple). Les individus sont numérotés à tout instant par $i \in I = \{1, 2, ..., N\}$. La population évolue à des pas de temps discrets, indicés par $n \in \mathbb{N}$. Plus précisément à chaque pas de temps, deux individus sont choisis uniformément au hasard, ils se reproduisent, et leur descendant remplace un troisième individu choisi indépendamment (mais pas nécessairement uniformément, plus de détails seront donnés juste après) dans la population. Notons qu'un individu peut éventuellement se reproduire avec lui-même mais cet événement n'a lieu qu'avec probabilité $1/N$ à chaque pas de temps. Sous cette dynamique la taille de population est constante, et à chaque pas de temps, au maximum 4 individus sont impliqués dans les changements de la population. Le fait que la composition de la population évolue très peu à chaque pas de temps nous permettra d'obtenir des résultats plus forts que ceux obtenus dans Derrida *et al.* (2000) pour le modèle de Wright-Fisher, dans lequel l'intégralité de la population est remplacée à chaque étape. Notons respectivement μ_n, π_n, $\kappa_n \in I$ les positions de la mère, du père, et de l'individu qui meurt au temps n. Les individus mère et père ont pour l'instant des rôles parfaitement symétriques.

Ajoutons de la sélection Le modèle présenté précédemment est dit neutre lorsque l'individu qui meurt au temps n est choisi uniformément dans la population. Une partie de notre travail, notamment les résultats les plus forts que nous obtenons, portent sur ce cas neutre. Néanmoins nous avons aussi étudié une version plus générale de ce modèle, qui permet de prendre en compte et d'étudier l'impact de la sélection génétique sur la composition génétique d'une population. Pour cela supposons maintenant que les individus peuvent être de deux types : avantagé ou non avantagé, et que cet avantage est conféré par une mutation qui se transmet de façon Mendélienne et haploïde. Plus simplement cela veut dire que lors d'une reproduction, l'un des deux parents, choisi uniformément au hasard, transmettra son statut d'avantagé ou non avantagé, à l'enfant produit. Comme les deux parents sont choisis uniformément au hasard dans la population, on supposera par convention que c'est la mère qui transmet son statut d'avantage. Enfin, cet avantage se manifeste au niveau de la mort des individus : on associe aux individus désavantagés un poids $1+s$ ($s \in \mathbb{R}_+ \cup \{+\infty\}$) et aux individus avantagés un poids 1, et au moment de choisir un individu qui meurt, chaque individu est choisi avec une probabilité proportionnelle à son poids. Ainsi, tant que le nombre d'avantagés est différent de 0 ou de N, un individu désavantagé a une probabilité de mourir $1+s$ fois plus élevée qu'un individu avantagé. Notons que lorsque s est infini, cela signifie que l'individu qui meurt est choisi uniformément parmi les individus désavantagés. Notons aussi que lorsque $s = 0$ le modèle est neutre, c'est-à-dire que tous les individus sont équivalents.

Transmission du génome et poids génétique des ancêtres Le processus stochastique $(\{\mu_n, \pi_n\}, \kappa_n)_{n\in\mathbb{N}}$ se traduit par un graphe de parenté, appelé pédigrée et noté G, qui est le support de la transmission génétique. Plus précisément, comme représenté dans la Figure 1.1, G est construit sur $I \times \mathbb{Z}_+$ en traçant à chaque temps $t \in \mathbb{Z}_+$ deux flèches orientées partant de $(\kappa_t, t+1)$ et allant vers (π_t, t) et (μ_t, t), et $N-1$ flèches orientées de $(x, t+1)$ à (x, t) pour tout $x \in I\backslash\{\kappa_t\}$. On notera $\{\mathcal{G}_n, n \in \mathbb{Z}_+\}$ la filtration associée au processus stochastique $(\{\mu_n, \pi_n\}, \kappa_n)_{n\in\mathbb{Z}_+}$.

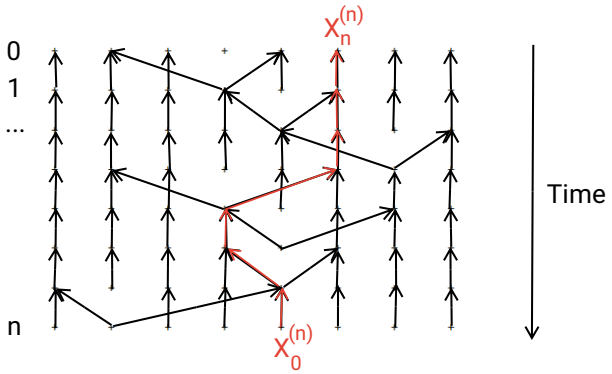


FIGURE 1.1 – Ce graphe représente le pédigrée d'une population de 8 individus, durant 7 pas de temps. Le chemin rouge représente la généalogie d'un gène, qui est la réalisation d'une marche aléatoire sur ce graphe, en remontant le temps.

Notons que le pédigrée n'est pas indépendant du statut d'avantage des individus et ne donne pas non plus cette information, qui est donnée par le processus stochastique complet $(\mu_n, \pi_n, \kappa_n)_{n\in\mathbb{N}}$ qui donne toute la dynamique de la population, à condition que les positions des individus initialement avantagés soient aussi connues. La Figure 1.2 donne un exemple de pédigrée dans lequel les individus avantagés sont en plus représentés en rouge.

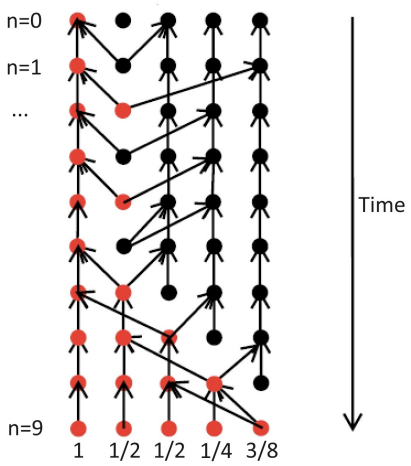


FIGURE 1.2 – Cette figure représente simultanément le pédigrée G et l'ensemble des individus avantagés (en rouge) durant 10 pas de temps, pour une population de 5 individus. Les nombres en bas donnent la probabilité pour qu'un gène échantillonné dans chacun des individus proviennent initialement de l'unique individu avantagé. Dans cet exemple le poids génétique de cet individu initialement avantagé vaut $21/8 = 1 + 1/2 + 1/2 + 1/4 + 3/8$.

Une fois que ce graphe est créé, si un gène (morceau de génome supposé insécable) est échantillonné dans un individu de la population, alors ce gène provient nécessairement de l'un de ses deux parents, choisi uniformément au hasard. La Figure 1.1 montre en rouge la généalogie, c'est-à-dire l'histoire d'un gène échantillonné au temps n dans l'individu 5. À ce stade il serait naturel de se pencher sur la question de la recombinaison, et de son impact sur le génome des individus. C'est une question difficile, qui est notamment abordée dans Lambert *et al.* (2018). De notre côté nous supposons plus simplement que le génome est constitué d'une infinité de loci (un locus est un emplacement du génome), qui se comportent de façon indépendante sur le pédigrée. Autrement dit, si l'on échantillonne cette fois deux gènes dans un individu, alors comme précédemment chacun de ces gènes provient de l'un des deux parents de l'individu, et l'on suppose ici ces provenances sont indépendantes, sachant ces deux parents. La généalogie de k gènes échantillonnés, c'est-à-dire l'histoire de k morceaux de génomes est donc un ensemble de k marches aléatoires sur le pédigrée, indépendantes sachant ce pédigrée. Sous cette hypothèse, le pédigrée étant donné, la probabilité pour qu'un gène échantillonné uniformément dans la population provienne d'un ancêtre donné peut être vue comme la proportion de génome issue de

cet ancêtre. C'est la quantité qui va nous intéresser. Plus formellement, notons $(X_k^{(n)}, n-k)_{0\leq k\leq n}$ la généalogie d'un gène (c'est-à-dire le numéro de l'individu dans lequel l'ancêtre de ce gène se trouvait à chaque temps $t = n - k \leq n$, dont un exemple est donné en Figure 1.1). Le processus $(X_k^{(n)}, n-k)_{0\leq k\leq n}$ est une marche aléatoire sur le graphe G et l'on considère

$$W_n(i,j) = \mathbb{P}(X_n^{(n)} = j | X_0^{(n)} = i, \mathcal{G}_n) \tag{1.1}$$

La quantité $W_n(i,j)$ modélise donc la proportion de génome de l'individu i qui provient de l'ancêtre j, ou encore la contribution de l'ancêtre j au génome de l'individu i. De la même façon, la quantité

$$M_n(j) = \sum_{i=1}^{N} W_n(i,j) \tag{1.2}$$

est égale à N fois la probabilité pour qu'un gène échantillonné uniformément au temps n provienne de l'ancêtre j, au temps 0. On appellera par la suite cette quantité le poids génétique de l'ancêtre j.

1.3 Résultats

Cas neutre Notre premier résultat porte sur le cas neutre, pour lequel $s = 0$. Le premier point du théorème dit que la contribution génétique d'un ancêtre donné est asymptotiquement la même dans tout individu vivant au temps présent. Le second point du théorème donne une loi explicite pour cette contribution asymptotique, lorsque la taille de population tend vers l'infini.

Théorème 1.1. (*i*) *Pour tout $j \in I$, il existe une variable aléatoire $A(j)$ telle que pour tout $i \in I$,*

$$W_n(i,j) \xrightarrow[n\to\infty]{} A(j) \quad p.s.$$

En particulier,

$$M_n(j) \xrightarrow[n\to\infty]{} M_\infty(j) = NA(j) \quad p.s.$$

(*ii*) *Pour tout $l \leq N$ et tous $k_1, ..., k_l \in \mathbb{Z}_+$,*

$$\mathbb{E}\left(M_\infty^{k_1}(1)...M_\infty^{k_l}(l)\right) \xrightarrow[N\to\infty]{} \prod_{i=1}^{l} 2^{k_i-1}k_i! \quad . \tag{1.3}$$

De façon équivalente, la contribution génétique asymptotique d'un ancêtre, est égale à 0 avec probabilité 1/2 (ce qui signifie que l'ancêtre en question a une descendance qui n'est pas éternelle), et sinon, suit une loi exponentielle de paramètre 1/2.

Quelques éléments de preuve de ce théorème et des résultats qui suivront (Théorèmes 1.2, 1.3, et Proposition 1.4) sont donnés dans la Section 1.4. Les deux points du théorème sont illustrés dans la Figure 1.3. Pour cette figure nous lançons une seule simulation du modèle

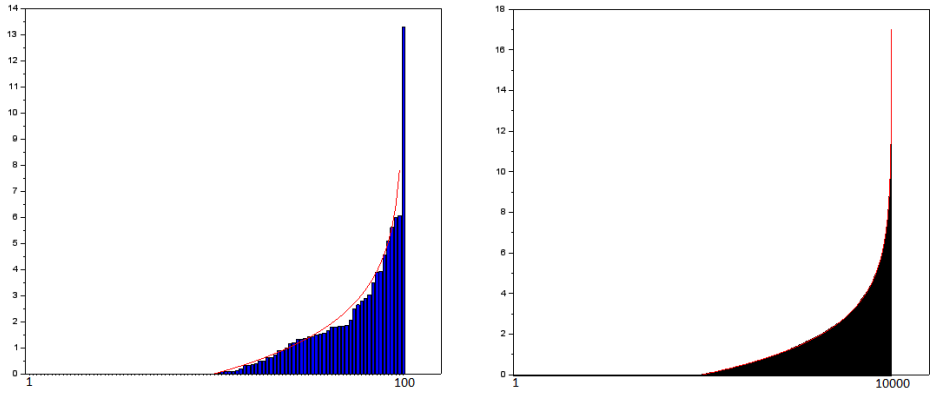


FIGURE 1.3 – Nous lançons une seule simulation du modèle de Moran biparental neutre, pour $N = 100$ (à gauche) ou $N = 10000$ (à droite) : en bleu nous traçons les poids des ancêtres triés dans l'ordre croissant, après 100000 pas de temps. En rouge nous donnons la fonction $x \mapsto -\mathbf{1}_{x>N/2} \times 2\ln(2(1-x/N))$.

de Moran biparental neutre, pour une durée assez longue. Nous obtenons que les poids des individus, triés dans l'ordre croissant, se stabilisent vers une distribution ; cette distribution converge vers l'inverse généralisée de la fonction de répartition de la variable aléatoire qui vaut 0 avec probabilité 1/2 et sinon suit une loi exponentielle de paramètre 1/2. Rappelons que dans le cas monoparental le comportement asymptotique du poids d'un ancêtre est très différent : pour toute taille de population N, le poids asymptotique d'un ancêtre est égal à 0 avec probabilité $(N-1)/N$ et sinon vaut N ; il existe un unique ancêtre commun à tous. Le Théorème 1.1 montre donc un comportement beaucoup plus complexe de la transmission génétique dans les populations à reproduction sexuée.

Notons que le poids asymptotique d'un ancêtre a une loi différente dans le modèle de Moran biparental et dans le modèle de Wright-Fisher biparental, étudié dans l'article Derrida *et al.* (2000). En particulier dans Derrida *et al.* (2000), les auteurs obtiennent que 80% des ancêtres ont une descendance éternelle. Cette différence est intéressante car ces modèles sont souvent considérés comme étant équivalents à changement d'échelle de temps près (pour ces deux modèles la dynamique de la proportion d'un allèle donné dans la population converge, une fois correctement renormalisée, vers une diffusion de Wright-Fisher). La différence que nous obtenons est liée au fait que lorsque la taille de population N est très grande, pour le modèle de Wright-Fisher biparental la loi du nombre d'enfants d'un individu est proche d'une loi de Poisson tandis que pour le modèle de Moran biparental elle est proche d'une loi géométrique à valeurs dans $\mathbb{N}$ (d'espérance 2 dans les deux cas). Ces deux distributions sont différentes et notamment la loi géométrique a une probabilité de valoir 0 beaucoup plus élevée (Figure 1.4).

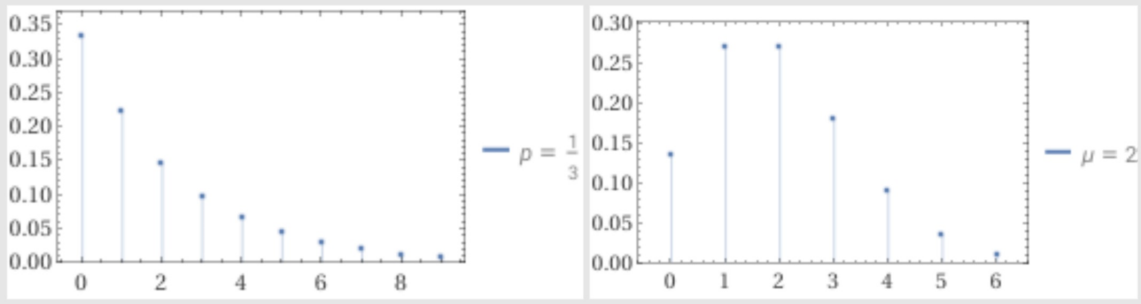


FIGURE 1.4 – Histogrammes de la loi géométrique (à gauche) et de la loi de Poisson (à droite) d'espérance 2.

Sélection infinie Revenons maintenant au cas avec sélection décrit dans la Section 1.2 et notons $\mathcal{Y}_n$ l'ensemble des individus avantagés au temps n et Y_n son cardinal, c'est-à-dire le nombre d'individus avantagés au temps n. Nous nous demandons dans ce cas, de façon assez naturelle, quelle est la probabilité pour qu'un gène échantillonné dans la population provienne des Y_0 individus initialement avantagés. Connaître cette probabilité permet de quantifier l'avantage conféré par une mutation avantageuse, en termes de quantité de génome transmis. En particulier on sait que cette probabilité sachant le pédigrée est une variable aléatoire (en tant que fonction déterministe du pédigrée qui est aléatoire), dont l'espérance vaut Y_0/N dans le cas neutre. Nous allons donc maintenant chercher à calculer cette espérance dans le cas avec sélection. Concentrons-nous pour l'instant sur le cas extrême où la force de la sélection est infinie ($s = +\infty$), et la population est initialement composée d'un seul individu avantagé. Le fait que s soit infini signifie qu'à chaque pas de temps, l'individu choisi pour mourir est choisi uniformément parmi les individus désavantagés. Dans ce cas, nous avons $Y_{n+1} \in \{Y_n, Y_n + 1\}$ pour tout n, donc le nombre d'individus avantagés croit avec le temps. Notons T_N le temps d'atteinte de N par $(Y_n)_{n\in\mathbb{N}}$, qui est fini presque sûrement. Alors le théorème suivant donne l'ordre de grandeur de la contribution génétique de l'individu initialement avantagé, une fois que tous les individus sont avantagés, c'est-à-dire au temps T_N. Notons que cette contribution continuera à évoluer après le temps T_N, de façon stochastique, mais son espérance restera la même car le modèle sera devenu neutre.

Théorème 1.2. *Le poids $M_{T_N}(1)$ de l'individu initialement avantagé (numéroté 1, par convention) au temps T_N satisfait*

$$\mathbb{E}(M_{T_N}(1)) \underset{N\to+\infty}{\sim} \frac{4}{\sqrt{\pi}}\sqrt{N}.$$

Ce théorème permet en quelque sorte de quantifier l'impact maximal de la sélection : en l'absence de sélection, la probabilité pour qu'un gène échantillonné uniformément dans la population provienne d'un ancêtre donné est égale à $1/N$. Lorsque la sélection est extrêmement forte, cette probabilité devient de l'ordre de $\frac{4}{\sqrt{\pi N}}$ pour l'individu initialement avantagé (elle reste donc en revanche de l'ordre de $1/N$ pour les autres individus). L'étude de la loi de ce poids asymptotique

est un projet en cours.

Sélection finie Lorsque la sélection n'est pas infinie, le nombre d'individus avantagés peut finir par toucher 0, avec une probabilité non nulle. Néanmoins cette probabilité tend vers 0 lorsque la taille de population tend vers l'infini et que la proportion d'individus avantagés est proche d'une constante positive $a \in (0,1)$. Nous nous sommes pour l'instant placés dans ce contexte. En travaillant sur le cas de sélection infinie nous avons compris que deux quantités importantes ont un comportement plutôt simple. Définissons pour tout temps $n \in \mathbb{N}$,

$$U_n = \sum_{l \in \mathcal{Y}_n} \sum_{l' \in \mathcal{Y}_0} W_n(l,l'), \quad \text{et} \quad V_n = \sum_{l \notin \mathcal{Y}_n} \sum_{l' \in \mathcal{Y}_0} W_n(l,l').$$

La quantité $U_n \in [0,N]$ (resp. $V_n \in [0,N]$) représente le poids génétique des individus initialement avantagés dans les individus avantagés (resp. désavantagés) au temps n. Tant que $Y_n \notin \{0,N\}$, si l'on note $\mathcal{U}(A)$ la loi uniforme sur un ensemble discret A, on a notamment

$$\frac{U_n}{Y_n} = \mathbb{P}(X_n^{(n)} \in \mathcal{Y}_0 | X_0^{(n)} \sim \mathcal{U}(\mathcal{Y}_n), \mathcal{G}_n),$$

et

$$\frac{V_n}{N - Y_n} = \mathbb{P}(X_n^{(n)} \in \mathcal{Y}_0 | X_0^{(n)} \sim \mathcal{U}(I \setminus \mathcal{Y}_n), \mathcal{G}_n).$$

Les deux quantités

$$\frac{U_{T_N}}{N}\mathbf{1}_{T_N<\infty} = \frac{U_{T_N}}{Y_{T_N}}\mathbf{1}_{T_N<\infty} \quad \text{et} \quad \frac{V_{T_0}}{N}\mathbf{1}_{T_0<\infty} = \frac{V_{T_0}}{N - Y_{T_0}}\mathbf{1}_{T_0<\infty}$$

peuvent donc être interprétées comme la contribution génétique des individus avantagés dans la population une fois que la mutation s'est fixée ou a disparu, respectivement. Elles donnent en effet la probabilité pour qu'un gène échantillonné uniformément dans la population devenue monomorphe, provienne d'un des individus initialement avantagés. La Figure 1.1 donne un exemple de telle probabilité, lorsque l'on part d'un seul individu avantagé : dans cette figure au bout de 9 pas de temps tous les individus sont avantagés, le poids de l'individu initialement avantagé vaut 21/8, donc la probabilité pour qu'un gène échantillonné uniformément au temps 9 provienne de cet individu vaut $21/8 \times 1/N = 21/40$. Pour étudier cette quantité limite U_{T_N}/N qui nous intéresse nous étudions le processus stochastique de dimension 3,

$$(Z_n)_{n\in\mathbb{N}} = \left(\frac{Y_n}{N}, \frac{U_n}{N}, \frac{V_n}{N}\right)_{n\in\mathbb{N}}.$$

Ce processus n'est pas Markovien, néanmoins lorsque la taille de population tend vers l'infini et la proportion initiale d'individus avantagés tend vers a, sa dynamique peut être approchée par celle d'un système dynamique dont la solution est explicite. C'est l'objet du prochain théorème.

Théorème 1.3. *Soit $a \in (0,1)$. Si la proportion initiale d'individus avantagés $\frac{Y_0}{N}$ tend vers a en probabilité quand N tend vers l'infini, alors pour tout $c \in \mathbb{R}_+$,*

$$\sup_{0\leq t\leq c} \|Z_{\lfloor Nt\rfloor} - z_t\| \underset{N\to\infty}{\longrightarrow} 0 \tag{1.4}$$

en probabilité, où $(z_t)_{t\geq 0} = (y_t, u_t, v_t)_{t\geq 0}$ satisfait

$$\begin{cases} y_t = F^{-1}\left(\dfrac{a^{1+s}}{1-a}\exp(st)\right) \qquad \text{où } F(x) = \dfrac{x^{1+s}}{1-x} \\ u_t = y_t \dfrac{a^{\frac{1+s}{2s}}}{(1-a)^{\frac{1}{2s}}}\left[\dfrac{(1-y_t)^{\frac{1}{2s}}}{y_t^{\frac{1+s}{2s}}} + \displaystyle\int_a^{y_t} \frac{(1-x)^{\frac{1}{2s}}}{x^{\frac{1+s}{2s}}}\left[\frac{1}{2} + \frac{1}{2s}\frac{1}{1-x}\right]dx\right] \\ v_t = (1-y_t)\dfrac{a^{\frac{1+s}{2s}}}{(1-a)^{\frac{1}{2s}}}\displaystyle\int_a^{y_t} \frac{(1-x)^{\frac{1}{2s}}}{x^{\frac{1+s}{2s}}}\left[\frac{1}{2} + \frac{1}{2s}\frac{1}{1-x}\right]dx. \end{cases} \tag{1.5}$$

Ce théorème donne la dynamique limite du triplet $(Z_n)_{n\in\mathbb{N}}$, et est en partie illustré dans la Figure 1.5. Notons que si la proportion initiale d'individus avantagés tend vers a, alors la probabilité pour que la mutation avantageuse se fixe, c'est-à-dire pour que $(Y_n)_{n\in\mathbb{N}}$ finisse par toucher N, tend vers 1. Notre résultat suivant étend alors le Théorème 1.3 jusqu'au temps d'intérêt T_N (qui tend vers l'infini lorsque N tend vers l'infini), pour l'espérance de $\frac{U_n}{N}$, c'est-à-dire la probabilité pour qu'un gène échantillonné dans la population provienne d'un des individus initialement avantagés.

Proposition 1.4. *Soit $a \in (0,1)$. Si la proportion initiale d'individus avantagés $\frac{Y_0}{N}$ tend vers a en probabilité quand N tend vers l'infini, alors*

$$\mathbb{E}\left(\frac{U_{T_N}}{N}\mathbf{1}_{T_N<\infty}\right) \underset{N\to\infty}{\longrightarrow} \frac{a^{\frac{1+s}{2s}}}{(1-a)^{\frac{1}{2s}}}\left(\int_a^1 \frac{(1-u)^{\frac{1}{2s}}}{u^{\frac{1+s}{2s}}}\left[\frac{1}{2} + \frac{1}{2s}\frac{1}{1-u}\right]du\right).$$

Notons que les deux processus stochastiques $(\frac{U_n}{N})_{n\in\mathbb{Z}_+}$ et $(\frac{V_n}{N})_{n\in\mathbb{Z}_+}$ continuent à évoluer après le temps $\inf(T_0, T_N)$ (leur dynamique sera détaillée dans la Section 1.4 donnant les éléments de preuves des résultats). Néanmoins leurs espérances respectives deviennent constantes après ce temps, car la population devient neutre. La Proposition 1.4 donne donc, sous une hypothèse de grande taille de population et proportion initiale macroscopique d'individus avantagés, l'espérance de la contribution génétique de ces individus avantagés (ou encore la probabilité pour qu'un gène échantillonné au temps présent provienne de l'un de ces individus). Lorsque la sélection est très forte ($s \to \infty$), la Proposition 1.4 nous dit que si la proportion initiale d'avantagés vaut a, alors ces individus avantagés seront en temps long à l'origine d'une proportion $2\sqrt{a} - a$ du génome de la population, en moyenne. A titre d'exemple, si initialement la population est constitué de 1% d'individus fortement avantagés, alors ces individus finiront par être responsables de 19% du génome de la population. Ce dernier résultat est illustré dans la Figure 1.5. La Figure 1.6 montre aussi l'impact maximal de la sélection, en traçant simplement les fonctions

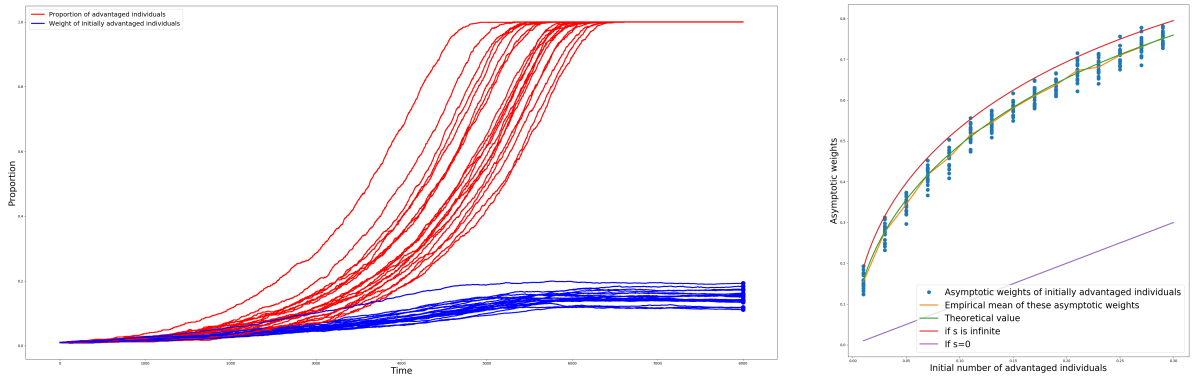


FIGURE 1.5 – Gauche : Pour $N = 1000$, $s = 10$ (sélection très forte), et $a = 1\%$, 20 réalisations de la dynamique jointe de la proportion d'individus avantagés (en rouge) et de la contribution génétique des individus initialement avantagés(en bleu). Droite : Pour différentes valeurs de la proportion initiale d'individus avantagés et $s = 10$ encore, 20 valeurs de la contribution génétique des individus initialement avantagés (points bleus), leur moyenne empirique (en jaune), l'approximation théorique de leur espérance (en bleu), ainsi que les approximation théorique de leur espérance pour $s = 0$ (en violet) et s infini (en rouge), une fois que la population est devenue monomorphe.

$x \to x$ et $x \to 2\sqrt{x} - x$ qui donnent les proportions initiales et finales de génome provenant d'une proportion x d'individus initialement très fortement avantagés. Pour finir, quelques éléments de comparaisons sont donnés dans la partie perspectives de ce chapitre (Section 1.5).

Pour finir, en partant d'une proportion a d'individus avantagés on perd le caractère aléatoire du poids asymptotique des individus avantagés, qui était bien caractérisé dans le cas neutre, par le Théorème 1.1. En particulier le poids des avantagés ne peut plus s'annuler en temps long alors qu'un seul individu a, dans le cas neutre, une probabilité égale à 1/2 d'avoir une descendance non éternelle. La proposition suivante donne la probabilité pour qu'un seul individu initialement avantagé (un mutant) ne contribue pas, en temps long à la population.

Proposition 1.5. *Supposons ici que le nombre initial d'avantagés est égal à 1 ($Y_0 = 1$). Alors la probabilité pour que le poids asymptotique de l'avantagé initial soit nul converge, lorsque N tend vers l'infini, vers*

$$\frac{\frac{3}{2} + \frac{1}{1+s} - \sqrt{\frac{9}{4} - \frac{s}{(1+s)^2}}}{2}.$$

Nous retrouvons bien sûr la valeur de 1/2 obtenue dans le Théorème 1.1, dans le cas où $s = 0$, et nous remarquons aussi que cette probabilité tend vers 0 lorsque s tend vers l'infini, ce qui est naturel. La Figure 1.7 donne la densité du poids d'un individu initialement très fortement avantagé, après un grand nombre de pas de temps. La proposition 1.5 renseigne sur la probabilité de l'atome en 0 de cette distribution. Le reste de la distribution dépend du nombre de pas de temps dans la simulation et fait l'objet de travaux en cours.

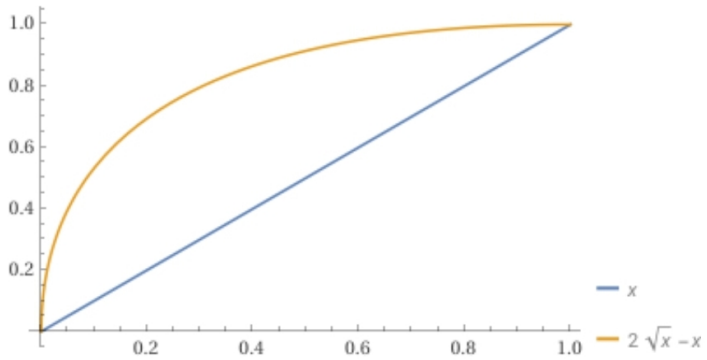


FIGURE 1.6 – Nous traçons les deux courbes $x \to x$ (en bleu) et $x \to 2\sqrt{x} - x$ (en orange). La différence entre les deux courbes montre l'impact maximal de la sélection, c'est-à-dire la différence entre les proportions initiales et finales de génome provenant d'une proportion x d'individus initialement très avantagés.

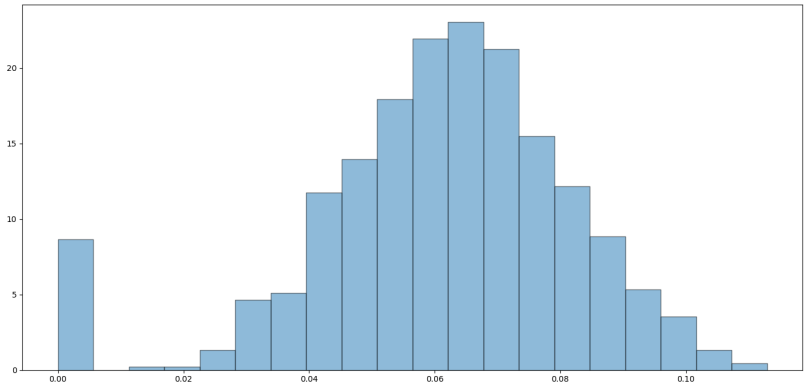


FIGURE 1.7 – Cette figure donne la densité du poids d'un ancêtre mutant fortement avantagé, pour $N = 10000$ et $s = 30$ (sélection très forte), au bout d'un temps long.

Bilan des résultats Nous avons défini un modèle de Moran biparental avec sélection et nous étudions pour ce modèle la contribution génétique asymptotique d'un individu donné, à l'ensemble de la population, c'est-à-dire la proportion du génome de la population qui provient de cet ancêtre, ou encore la probabilité pour qu'un gène échantillonné au temps présent provienne de cet ancêtre. Dans le cas le plus simple, c'est-à-dire en l'absence de sélection, nous montrons que cette probabilité, multipliée par N, converge en loi lorsque la taille de population N tend vers l'infini, vers une variable aléatoire dont la loi est explicite : elle vaut 0 avec probabilité $1/2$ et sinon suit une loi exponentielle de paramètre $1/2$. Dans un cas de sélection infinie le nombre d'avantagés croit avec le temps, et nous montrons que l'espérance de la proportion de génome transmise à la population par un unique individu initialement avantagé, est de l'ordre de $\frac{4}{\sqrt{\pi N}}$ (contre $1/N$ dans le cas neutre). Dans le cas le plus général de sélection finie, le nombre d'avantagés ne croit plus avec le temps, mais à condition de supposer que la proportion initiale d'avantagés, a, est strictement positive, nous prouvons que l'espérance de la proportion de génome transmise à la population par l'ensemble des individus initialement avantagés converge lorsque la taille de population N tend vers l'infini, vers

$$\frac{a^{\frac{1+s}{2s}}}{(1-a)^{\frac{1}{2s}}}\left(\int_a^1 \frac{(1-u)^{\frac{1}{2s}}}{u^{\frac{1+s}{2s}}}\left[\frac{1}{2}+\frac{1}{2s}\frac{1}{1-u}\right]du\right)$$

(contre a dans le cas neutre). Lorsque s tend vers l'infini, cette quantité tend vers $2\sqrt{a}-a$. Cette quantification de l'impact de la sélection sur la proportion de génome transmise est illustré dans la Figure 1.5 (figure de droite).

1.4 Éléments de preuves

Éléments de preuve du Théorème 1.1

Équation stationnaire pour déterminer la loi limite Pour démontrer le Théorème 1.1 nous commençons par trouver la loi asymptotique du poids d'un ancêtre. En effet, lorsque N tend vers l'infini, cette loi doit satisfaire une équation stationnaire qui peut être résolue (cette stationnarité sera perdue dans le cas avec sélection). Plus précisément notons $h(\lambda) = \mathbb{E}(\exp(-\lambda M_\infty(1)))$ la transformée de Laplace du poids asymptotique de l'individu 1 (qui ne dépend pas du numéro de cet individu, puisque l'on se place dans le cas neutre pour ce premier théorème). En supposant (sans le démontrer dans un premier temps) l'indépendance des poids asymptotiques des ancêtres, nous trouvons que cette transformée de Laplace doit satisfaire l'équation

$$h(\lambda) = \frac{1}{3} + \frac{2}{3}h\left(\frac{\lambda}{2}\right)h(\lambda).$$

En cherchant une solution à cette équation sous la forme $(1+a\lambda)/(1+b\lambda)$ nous obtenons que ce poids asymptotique a une loi simple : il est soit égal à 0 avec probabilité $1/2$, soit suit une loi exponentielle de paramètre $1/2$.

Marches aléatoires indépendantes sur le pédigrée Ce premier élément de preuve nous permet de déterminer les moments asymptotiques donnés dans le terme de droite de l'Équation (1.3). La suite de la preuve du Théorème consiste alors notamment à prouver la convergence des moments

$$\mathbb{E}\left(M_\infty^{k_1}(1)...M_\infty^{k_l}(l)\right)$$

vers ces valeurs asymptotiques lorsque la taille de population tend vers l'infini. Pour cela, nous posons $k = \sum_{j=1}^{l} k_j$, et nous introduisons k marches aléatoires indépendantes sur le pédigrée G, qui partent de positions uniformes et indépendantes sur I au temps n et remontent le temps : $(X_i^{(1,n)}, n-i)_{i\leq n}$, $(X_i^{(2,n)}, n-i)_{i\leq n}$, ..., $(X_i^{(k,n)}, n-i)_{i\leq n}$. Alors par définition de $M_n(j)$ (Équations (1.1) et (1.2)), pour tous $k_1, k_2, ..., k_l \in \mathbb{Z}_+$ tels que $\sum_{j=1}^{l} k_j = k$, on a

$$\mathbb{E}\left(M_n^{k_1}(1)...M_n^{k_l}(l)\right) = N^k\,\mathbb{P}(X_n^{(1,n)} = ... = X_n^{(k_1,n)} = 1, ..., X_n^{(k_1+...+k_{l-1}+1,n)} = ... = X_n^{(k_1+...+k_l,n)} = l).$$

En faisant tendre n vers l'infini et en notant $\nu_{N,k}$ la loi stationnaire de la chaîne de Markov $\left(X_n^{(1,n)}, X_n^{(2,n)}, ..., X_n^{(k,n)}\right)_{n\in\mathbb{N}}$ qui est irréductible et apériodique, à valeurs dans I^k, on a alors que pour tous $k_1, k_2, ..., k_l \in \mathbb{Z}_+$ tels que $\sum_{j=1}^{l} k_j = k$,

$$\mathbb{E}\left(M_\infty^{k_1}(1)...M_\infty^{k_l}(l)\right) = N^k \nu_{N,k}(1, ..., 1, 2, ..., 2, ..., l, ..., l)$$

où dans le terme de droite, le nombre $j \in \{1, 2, ..., l\}$ est répété k_j fois. Il reste alors à prouver que

$$N^k \nu_{N,k}(1, ..., 1, 2, ..., 2, ..., l, ..., l) \xrightarrow[N\to+\infty]{} \prod_{i=1}^{l} 2^{k_i-1} k_i!. \tag{1.6}$$

La loi stationnaire $\nu_{N,k}$ de $(X_n^{(1,n)}, ..., X_n^{(k,n)})_{n\in\mathbb{N}}$ est l'unique mesure de probabilité solution de l'équation

$$\nu_{N,k} = \nu_{N,k} Q^{(N,k)} \tag{1.7}$$

où $Q^{(N,k)}$ est la matrice de transition de la chaîne de Markov $(X_n^{(1,n)}, ..., X_n^{(k,n)})_{n\in\mathbb{N}}$. Nous cherchons donc à montrer d'une part que pour tout $x \in I^k$, $\nu_{N,k}(x)$ est équivalent à $C(x)N^{-k}$, et d'autre part que $C(x)$ a la forme appropriée, c'est-à-dire que $C(x) = \prod_{i=1}^{N} 2^{K_i(x)}$ où $K_i(x)$ est le nombre d'occurrences du nombre i dans le vecteur x. Nous montrons ces deux points l'un après l'autre.

Projection Le premier point est démontré en introduisant une projection assez naturelle (nous verrons pourquoi juste après), notée $(Y_n^{(k)})_{n\in\mathbb{N}}$, de la chaîne de Markov $(X_n^{(1,n)}, ..., X_n^{(k,n)})_{n\in\mathbb{N}}$, sur un sous-espace de I^k. Plus précisément, pour tout $x = (x_1, ..., x_k) \in I^k$, définissons la *configuration associée à* x, comme l'ensemble $\{x\} = \{k_1, k_2, ..., k_l\}$ des nombres de répétitions de chaque élément de I présent dans x. Le nombre l, aussi noté $L(\{x\})$, est appelé la taille de la configuration

{x}. Par exemple, si $N \geq 4$, $k = 4$ et et $x = (3, 1, 4, 4)$ alors $\{x\} = \{1, 1, 2\}$ et donc $L(\{x\}) = 3$. Nous posons alors pour tout $n \geq 0$, $Y_n^{(k)}$ la configuration associée à $(X_n^{(1,n)}, ..., X_n^{(k,n)})$. Cette projection est naturelle du fait de l'invariance de la loi de la chaîne de Markov $(X_n^{(1,n)}, ..., X_n^{(k,n)})_{n\in\mathbb{N}}$ d'une part par changement de numérotation des N sites et d'autre part par changement de numérotation des k particules. Grâce à cette invariance, la projection $(Y_n^{(k)})_{n\in\mathbb{N}}$ est en effet encore une chaîne de Markov, et

$$\nu_{N,k}(x) \sim \frac{\nu_{N,k}(\{x\})}{N^l} \frac{\prod_{i=1}^{l} k_i!}{k!} \times \prod_{j=1}^{k} n_j!, \tag{1.8}$$

où $\nu_{N,k}(\{x\}) = \lim_{n\to\infty} \mathbb{P}(Y_n^{(k)} = \{x\})$ est la probabilité asymptotique que la chaîne de Markov $(X_n^{(1,n)}, ..., X_n^{(k,n)})_{n\in\mathbb{N}}$ soit dans la configuration $\{x\}$.

Distribution stationnaire de la chaîne de Markov projetée La fin de la preuve consiste à étudier cette nouvelle quantité $\nu_{N,k}(\{x\})$. La chaîne de Markov $(Y_n^{(k)})_{n\in\mathbb{N}}$ prend ses valeurs dans l'espace

$$\mathcal{S}_k = \{y = \{k_1, ..., k_l\} | k_j \in \mathbb{N}^* \, \forall j, \sum_{i=1}^{l} k_i = k\}$$

qui ne dépend plus de N. A chaque pas de temps, la chaîne de Markov $(Y_n^{(k)})_{n\in\mathbb{N}}$ peut, partant de l'état $\{k_1, ..., k_l\}$, soit rester dans le même point, soit sauter dans une autre configuration, qui aura pour taille $l-1$ (si $l \geq 2$), l, ou $l+1$ (si $l \leq k-1$). Les probabilités de transition d'un état à l'autre sont d'un ordre de grandeur qui dépend de la différence de taille entre les configurations initiale et finale, comme représenté dans la Figure 1.8. En particulier la probabilité pour que la taille de configuration diminue ou reste sur place est de l'ordre de C/N^2 où C dépend de l'état de départ, tandis que la probabilité pour que la taille de configuration augmente est d'ordre C/N où C dépend de l'état de départ (une façon d'interpréter ce résultat est de remarquer que lorsque la taille de population est grande, les marches aléatoires sur le pédigrée ont tendance à se trouver sur des sites différents : elles coalescent à un taux beaucoup plus faible qu'elles ne se séparent). Notre preuve s'appuie pour finir sur la caractérisation des distributions stationnaires donnée dans Shubert (1975). Plus précisément, pour toute configuration $y \in \mathcal{S}_k$ introduisons l'ensemble $G(y)$ des arbres couvrants orientés enracinés et dirigés vers y, qui sont inclus dans le graphe de transition de $Y^{(k)}$. Pour tout arbre orienté $g \in G(y)$, définissons son poids $\pi(g)$ comme le produit des probabilités de ses arêtes, pour la chaîne de Markov $Y^{(k)}$. Alors d'après Shubert (1975), la distribution stationnaire de la chaîne de Markov $Y^{(k)}$ est telle que pour tout $y \in S_k$,

$$\nu_{N,k}(y) = \frac{\sum_{g\in G(y)} \pi(g)}{\sum_{y'\in S_k} \sum_{g'\in G(y')} \pi(g')}. \tag{1.9}$$

Maintenant d'après les équivalents des probabilités de transition évoqués précédemment, la probabilité $\pi(g)$ d'un arbre couvrant orienté g pointé vers y est d'ordre au maximum

$$\frac{C(\mathcal{T})}{N^{2(k-l)}} \frac{1}{N^{\#S_k - 1 - (k-l)}} = \frac{C(\mathcal{T})}{N^{\#S_k - 1 + (k-l)}}, \tag{1.10}$$

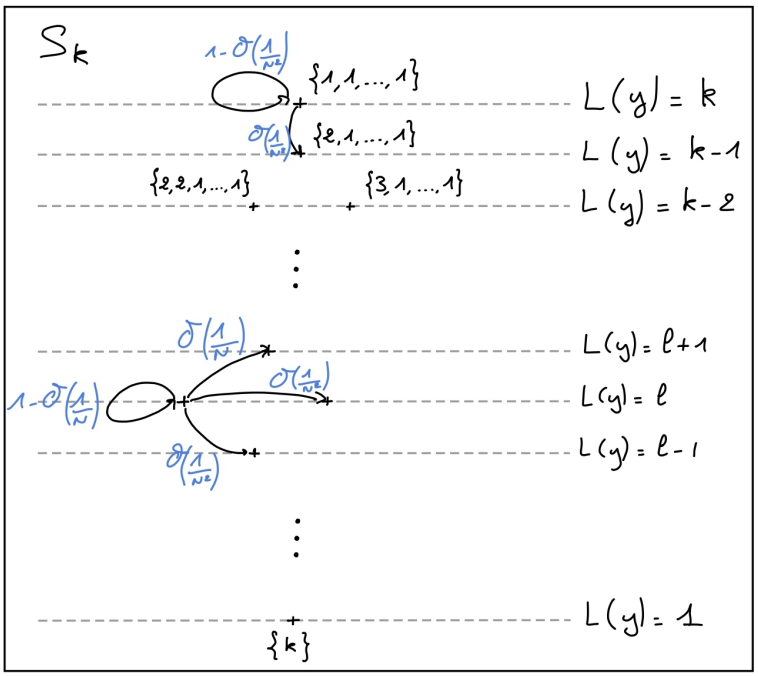


FIGURE 1.8 – Représentation schématique de l'espace d'états et des probabilités de transition de la chaîne de Markov $Y^{(k)}$. Les états sont rangés de haut en bas en fonction de leur taille : l'état le plus haut correspond au cas où les k particules sont dans des états différents, tandis que l'état le plus bas correspond au cas où toutes les particules sont dans le même site.

où la quantité $C(\mathcal{T})$ ne dépend pas de N, et il existe un tel arbre ayant effectivement une probabilité de cet ordre. Cet arbre peut être construit comme représenté dans la Figure 1.9, en commençant par tracer un chemin strictement descendant (au sens où la taille de la configuration décroît strictement le long de ce chemin), allant de $\{1,1,...,1\}$ à y puis en y ajoutant des arêtes strictement ascendantes partant de chaque configuration qui n'est pas sur ce chemin. En combinant le calcul (1.10) avec l'Équation (1.9), on obtient que

$$\nu_{N,k}(\{x\}) \sim C(\{x\})\frac{N^{L(\{x\})}}{N^k} \qquad \text{quand } N \text{ tend vers l'infini.} \tag{1.11}$$

Revenons enfin à la chaîne de Markov qui nous intéresse, $(X_n^{(1,n)},...,X_n^{(k,n)})_{n\in\mathbb{N}}$. Les équations (1.8) et (1.11) nous donnent que sa loi stationnaire satisfait pour tout $x \in I^k$:

$$\nu_{N,k}(x) \sim \frac{K(\{x\})}{N^k} \qquad \text{quand } N \text{ tend vers l'infini,} \tag{1.12}$$

où $K(\{x\}) = C(\{x\})\frac{\prod_{i=1}^{l} k_i!}{k!} \times \prod_{j=1}^{k} n_j!$ ne dépend pas de N.

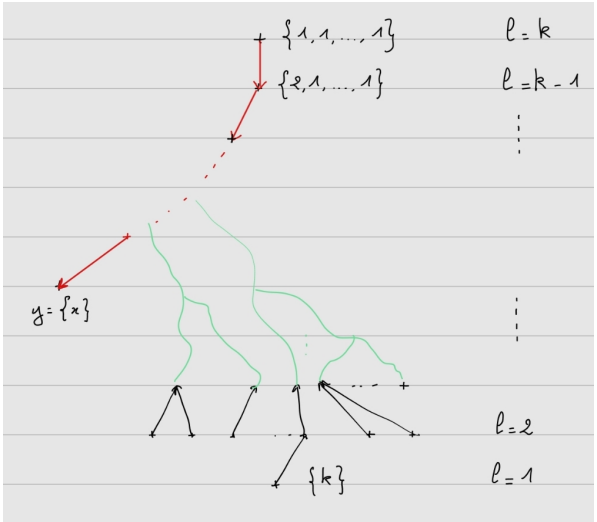


FIGURE 1.9 – Représentation schématique d'un arbre couvrant pointé vers la configuration $y = \{x\}$, dont la probabilité est d'ordre maximal.

Il reste à prouver que $K(\{x\}) = \prod_{i=1}^{l} 2^{k_i-1} k_i!$. Pour cela, rappelons que la loi stationnaire $\nu_{N,k}$ de $(X_n^{(1,n)}, ..., X_n^{(k,n)})_{n\in\mathbb{N}}$ est l'unique mesure de probabilité solution de

$$\nu_{N,k} = \nu_{N,k} Q^{(N,k)}.$$

En prenant le premier ordre (connu, grâce à l'Équation (1.12)) de cette équation, on obtient que K est solution de

$$K(\{k_1, ..., k_l\}) \times \left[l - 2 \sum_{\mu=1}^{l} \left(\frac{1}{2}\right)^{k_\mu} \right] = 2 \sum_{\mu=1}^{l} \sum_{i=1}^{k_\mu - 1} \left(\frac{1}{2}\right)^{i} \binom{k_\mu}{i} K(\{k_1, ..., k_\mu - i, ..., k_l, i\}). \tag{1.13}$$

Nous concluons la preuve du Théorème 1.1 en montrant que les solutions de cette équation (qui peut être vue comme une équation de récurrence sur la taille des configurations) sont toutes proportionnelles, puis que pour toute constante C, $K(\{x\}) = C \times \prod_{i=1}^{l} 2^{k_i-1} k_i!$ est une solution de cette équation, et enfin que cette constante C vaut nécessairement 1 si l'on veut que $\nu_{N,k}$ soit une mesure de probabilité.

Éléments de preuve du Théorème 1.2 et de la Proposition 1.4

Lorsque l'on ajoute de la sélection, la population est constitué à chaque instant de Y_n individus avantagés, et ce nombre Y_n est une chaîne de Markov. Dans le cas où la sélection est infinie, cette chaîne de Markov peut à chaque pas de temps soit augmenter de 1, soit rester sur place. Elle est par ailleurs absorbée en N. Le Théorème 1.2 donne l'ordre de grandeur de l'espérance du

poids limite de l'unique individu initialement avantagé, donc l'ordre de grandeur de la probabilité asymptotique pour qu'un gène échantillonné dans la population provienne de cet individu. Pour obtenir ce résultat nous supposons que l'individu initialement avantagé est l'individu 1, par convention, et nous définissons deux quantités assez naturelles, présentées aussi dans la Section 1.3 :

$$U_n = \sum_{l \in \mathcal{Y}_n} W_n(l,1), \quad \text{et} \quad V_n = \sum_{l \notin \mathcal{Y}_n} W_n(l,1).$$

Les quantités U_n et V_n donnent les poids génétiques de l'individu initialement avantagé, respectivement dans les individus avantagés et les individus désavantagés au temps n. Ces quantités sont naturelles car elles consistent à regarder le poids de chaque individu dans chaque autre individu, en les regroupant par statut d'avantage (les individus ayant le même statut à un instant donné étant échangeables).

La preuve du Théorème 1.2 consiste à considérer que la trajectoire de $(Y_n)_{n\in\mathbb{N}}$ est connue, et à regarder la dynamique de la population, sachant cette réalisation de trajectoire. Concrètement, en notant $(\mathcal{F}_n, n \in \mathbb{N})$ la filtration associée à la chaîne de Markov Y, le modèle défini dans la Section 1.2 nous donne, lorsque le paramètre de sélection s est infini, que la dynamique de $(U_n, V_n)_{n\in\mathbb{N}}$ satisfait

$$\begin{aligned}
\mathbb{E}(U_{n+1}|Y_{n+1} = Y_n, \mathcal{F}_n) &= U_n, \\
\mathbb{E}(U_{n+1}|Y_{n+1} = Y_n + 1, \mathcal{F}_n) &= U_n + \frac{U_n}{2Y_n} + \frac{1}{2N}(U_n + V_n), \\
\mathbb{E}(V_{n+1}|Y_{n+1} = Y_n, \mathcal{F}_n) &= V_n - \frac{V_n}{N - Y_n} + \frac{V_n}{2(N - Y_n)} + \frac{1}{2}\frac{U_n + V_n}{N}, \\
\mathbb{E}(V_{n+1}|Y_{n+1} = Y_n + 1, \mathcal{F}_n) &= V_n - \frac{V_n}{N - Y_n}.
\end{aligned}$$

Rappelons maintenant que la chaîne de Markov $(Y_n)_{n\in\mathbb{N}}$ part de 1 puis pour tout $k \in \mathbb{N}^*$ saute de k à $k+1$ au bout d'un temps qui suit une loi géométrique (à valeurs dans $\mathbb{N}^*$) dont le paramètre dépend de k. Pour tout $k \in \{1, ..., N\}$, notons $S_k = \inf\{n \in \mathbb{N} | Y_n = k\}$, $u_k = \mathbb{E}(U_{S_k})$, et $v_k = \mathbb{E}(V_{S_k})$. Grâce à la dynamique simple de $(Y_n)_{n\in\mathbb{N}}$, on peut montrer que la suite (finie) de vecteurs $\begin{pmatrix} u_k \\ v_k \end{pmatrix}_{k\in\{1,...,N\}}$ satisfait pour tout $k \in \{1, ..., N-1\}$

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \mathcal{A}_k \begin{pmatrix} u_k \\ v_k \end{pmatrix} \tag{1.14}$$

où

$$\mathcal{A}_k = \sum_{l=0}^{+\infty} L_k (H_k)^l = L_k [I - H_k]^{-1}$$

(le nombre l représente ici le nombre de fois où la chaîne de Markov Y est restée au niveau k avant de passer au niveau $k+1$) et

$$H_k = \left(1 - \frac{k}{N}\right) \begin{pmatrix} 1 & 0 \\ \frac{1}{2N} & 1 - \frac{1}{2(N-k)} + \frac{1}{2N} \end{pmatrix},$$

$$L_k = \frac{k}{N} \begin{pmatrix} 1 + \frac{1}{2k} + \frac{1}{2N} & \frac{1}{2N} \\ 0 & 1 - \frac{1}{N-k} \end{pmatrix}.$$

La fin de la preuve consiste à observer alors que si $\tilde{u}_k = \frac{u_k}{k}$ et $\tilde{v}_k = \frac{v_k}{N-k}$ pour tout $k \in \{1, ..., N-1\}$, alors

$$\tilde{u}_{k+1} - \tilde{v}_{k+1} = \frac{2Nk + N + k}{(2N+1)(k+1)}(\tilde{u}_k - \tilde{v}_k) \tag{1.15}$$

pour tout $k+1 < N$. Ceci nous donne une expression de la différence $\tilde{u}_k - \tilde{v}_k$ comme un produit que nous pouvons contrôler :

$$\frac{2}{\sqrt{\pi k}}\left(1 - \frac{C}{k} - \frac{C(\log(k)+1)}{N}\right) \leq x_k \leq \frac{2}{\sqrt{\pi k}}\left(1 + \frac{C}{k}\right).$$

En observant finalement, grâce à l'Équation (1.14) à nouveau, que pour tout $k \in \{1, ..., N-1\}$,

$$\tilde{v}_{k+1} = \frac{x_k}{2N+1} + \tilde{v}_k$$

nous pouvons alors donner un équivalent de v_N et donc de u_N lorsque N tend vers l'infini.

La Proposition 1.4 (correspondant au cas où le paramètre de sélection s est fini) est prouvée de la même façon, sauf que les équations de récurrences obtenues sont différentes, et nous utilisons aussi des arguments de monotonie des espérances de U_n/Y_n et $V_n/(N-Y_n)$ en fonction de n. Le détail de la preuve est donné dans la note Coron et Le Jan (2024c), Théorème 2.9).

Éléments de preuve du Théorème 1.3

Le Théorème 1.3 dit que le processus stochastique $Z_n = (Y_n/N, U_n/N, V_n/N)_{n\in\mathbb{N}}$ qui prend ses valeurs dans $[0,1]^3$ et est constitué de la proportion d'avantagés dans la population au temps n, Y_n/N, et du poids génétique des avantagés initiaux parmi les avantagés (U_n/N) et les désavantagés (V_n/N) de la population, converge vers une unique solution d'un système dynamique. Ce processus $(Z_n)_{n\in\mathbb{N}}$ part de l'état $(a, a, 0)$, où $a \in (0,1)$ est la proportion initiale d'avantagés dans la population. Pour démontrer le Théorème 1.3 nous étudions la dynamique de ce processus stochastique. Tout d'abord, comme dans le cas à sélection infinie, le processus stochastique $(Y_n)_{n\in\mathbb{N}}$ est une chaîne de Markov, et est même assez simple : il s'agit du changement de temps aléatoire d'une marche aléatoire simple sur $\{0, 1, ..., N\}$, absorbée en 0 et N. En notant $(\mathcal{H}_n)_{n\in\mathbb{N}}$ la filtration associée au processus $(Z_n)_{n\in\mathbb{N}}$, on obtient par ailleurs que lorsque la taille de population N tend vers l'infini,

$$\mathbb{E}\left(Z_{n+1} - Z_n | \mathcal{H}_n\right) = \frac{1}{N} g(Z_n) + o\left(\frac{1}{N}\right),$$

où

$$g(y,u,v) = \Big(\frac{y(1-y)s}{y+(1+s)(1-y)},\\ \frac{u}{2}+\frac{u+v}{2}y-\frac{u}{y+(1+s)(1-y)},\\ \frac{v}{2}+\frac{u+v}{2}(1-y)-\frac{(1+s)v}{y+(1+s)(1-y)}\Big)$$

pour tout $(y,u,v)\in[0,1]^3$. Ce calcul nous indique la forme du système dynamique limite, dont nous déterminons la forme de l'unique solution partant de $(a,a,0)$ (donnée dans l'Équation (1.5)), notamment en utilisant la fonction $(y_t, t\geq 0)$ comme un changement de temps. Nous obtenons alors par des techniques assez classiques la convergence du processus stochastique $(Z_{\lfloor Nt\rfloor})_{t\leq c}$ vers cette solution jusqu'à tout temps fini c.

1.5 Perspectives

Ce travail de recherche a de nombreuses perspectives et a pour commencer été poursuivi par les encadrements de Luce Breuil (stage de 3e année de l'École Polytechnique) d'une part, et Raphaël Tran Thanh et Juan Mardomingo Sanz (projet de Master 2 Mathématiques pour les sciences du vivant) d'autre part. Luce Breuil a montré la généralisation du Théorème 1.1 dans le cas où les individus ont m parents, avec m entier naturel fixé supérieur à 2. Elle a aussi étudié le poids asymptotique des ancêtres dans le cas où les individus peuvent soit se reproduire avec eux-mêmes (auto-fécondation) avec une certaine probabilité p, soit se reproduire avec un autre individu (allo-fécondation). Pour ce travail de stage Luce a obtenu le prix de stage de la chaire Modélisation Mathématique et Biodiversité. L'impact de l'auto-fécondation sur la composition génétique d'une population est aussi étudié dans l'article très récent Newman *et al.* (2024), qui cherche aussi à prendre en compte le rôle du pédigrée dans cette question, et j'aimerais comprendre les liens entre les résultats obtenus dans cet article et ceux que nous pouvons fournir en utilisant l'approche que nous avons empruntée, avec Yves Le Jan d'une part et avec Luce Breuil d'autre part. Raphaël Tran Thanh et Juan Mardomingo Sanz étudient actuellement une situation dans laquelle la sélection n'a pas lieu lors de la mort mais lors de la reproduction. Plus précisément ils supposent que la population est constituée de deux types d'individus, et que le type d'un individu influence d'une part sa capacité de reproduction, et d'autre part le choix du type de son partenaire. La question est alors de déterminer dans quelle mesure ces préférences d'appariement influencent la contribution génétique d'un ancêtre donné et aussi de comparer les forces de la sélection naturelle et de la sélection par préférences d'appariement, ce qui peut être fait notamment en prenant un paramètre de sélection infini. Ils obtiennent en effet des résultats quantitatifs similaires à celui que nous obtenons dans le Théorème 1.2 : lorsque la sélection implique que l'un des deux parents est nécessairement du type avantagé, et si la population est initialement constituée de 1% d'individus avantagés, alors ceux-ci seront en temps long à

l'origine de 14% du génome total de la population. Je co-encadre ce projet avec Diala Abu Awad (Université Paris-Saclay) qui est notamment spécialiste du rôle des systèmes de reproduction dans l'évolution génétique des populations et avec laquelle j'ai déjà étudié des modèles probabilistes de démo-génétique des populations et le rôle des traits d'histoire de vie dans l'évolution génétique (Abu Awad et Coron (2018)). Enfin ce sujet de recherche qui repose sur un modèle probabiliste très simple peut aussi être relié à d'autres approches d'étude de la diversité génétique qui sont beaucoup plus appliquées, notamment les coalescents séquentiellement Markoviens et l'estimation de paramètres démographiques à partir de données génétiques. Sur des thématiques proches je vais prochainement co-encadrer, avec Paul Verdu (Musée de l'Homme) et Tristan Mary-Huard (INRAE), le stage de Master 2 et la thèse de Gaspard Dousson-Lys, qui portera sur l'estimation de paramètres d'histoires migratoires, et l'estimation de paramètres de sélection naturelle, à partir de données de mesure d'hybridation dans le génome d'individus échantillonnés dans une population qui résulte des contributions successives de populations sources. Je co-encadre déjà le projet de master 2 de Gaspard Dousson-Lys et Angelo Ciambelli sur des questions liées.

Chapitre 2

Préférences d'appariement : évolution et rôle dans la diversité génétique

Le chapitre précédent portait sur l'étude de la proportion du matériel génétique d'une population à reproduction sexuée, qui est issue d'un ancêtre donné. Nous avons en particulier étudié l'impact de la sélection sur cette proportion et j'ai mentionné comme perspective de recherche l'étude de l'impact des préférences d'appariement sur la composition génétique d'une population à reproduction sexuée. Cette dernière question rejoint les sujets d'un ensemble de travaux de recherche que j'ai menés avec Manon Costa (Institut de Mathématiques de Toulouse), Hélène Leman (INRIA, Lyon) et Charline Smadi (INRAE, Grenoble). Ce chapitre porte sur l'étude, à partir de modèles individus centrés, de l'évolution des préférences d'appariement et de leur rôle sur la composition génétique des populations.

2.1 Introduction

La sélection naturelle, introduite dans Darwin (1859), repose sur les mutations du génome, qui, de façon aléatoire, apparaissent et sont transmises par un individu à son descendant. Certaines de ces mutations n'ont pas d'impact sur la survie ou la reproduction de leur porteur, tandis que d'autres peuvent lui conférer un avantage, ou un désavantage. Les versions mutées du génome qui donnent un avantage à leur porteur ont alors plus de chances d'être transmises et sont ainsi sélectionnées, au travers du comportement des individus (Dawkins (1976)). Ce mélange de mutations avantageuses et neutres permet une diversité génétique au sein des populations, et aussi une adaptation progressive et différenciée des populations à différents environnements. Ces adaptations différenciées peuvent aller jusqu'à un arrêt des flux de gènes entre des populations adaptées à des environnements différents : c'est ce que l'on appelle la spéciation écologique. La spéciation est un processus évolutif complexe au cours duquel deux groupes d'individus d'une même espèce finissent par ne plus pouvoir se reproduire ensemble, et donc finir par appartenir à deux espèces différentes. La modélisation et l'étude mathématique de la spéciation est un su-

jet important qui a connu des progrès récents, notamment au travers des travaux Couvert *et al.* (2024); Bard (2023) qui abordent cette question en développant des modèles multi-échelles, allant du gène à l'espèce, en passant par la protéine et l'individu.

La question principale qui nous a intéressées est celle du rôle de la reproduction sexuée et des préférences d'appariement, d'une part dans la spéciation, et d'autre part dans le maintien de la diversité génétique des populations. Ces sujets font aussi l'objet de travaux récents réalisés par des généticiens des populations (Marie-Orleach *et al.* 2024; Shaw *et al.* 2024). Le point de départ de notre travail a été l'étude de l'article M'Gonigle *et al.* (2012) étudiant la spéciation par sélection sexuelle, c'est-à-dire l'arrêt de flux de gènes entre deux sous-populations, du fait de préférences d'appariement. Nous nous sommes alors penchées sur deux types de préférences d'appariement : l'homogamie (le fait pour un individu de préférer se reproduire avec un individu similaire, ou du même type) et l'hétérogamie (le fait au contraire de préférer se reproduire avec un individu de phénotype ou génotype différent). Nous avons étudié trois questions :

- Sous quelles conditions l'homogamie peut représenter un avantage sélectif et finir par envahir une population (Coron *et al.* (2021), en collaboration avec Fabien Laroche, d'INRAE)
- Comment l'homogamie couplée à une structure spatiale peut conduire à de la spéciation (Coron *et al.* (2018b))
- Quel niveau de diversité génétique est généré par l'hétérogamie (Coron *et al.* (2022)), en collaboration avec Violaine Llaurens, du Museum National d'Histoire Naturelle).

2.2 Modèle

Dans le chapitre précédent j'ai présenté et étudié un modèle de Moran, qui est un modèle dit "de population", car les paramètres qui caractérisent ce modèle, en l'occurrence essentiellement la taille N de population, s'interprètent à l'échelle de la population. Dans ce chapitre, nous considérons des modèles dits "individu-centrés", dont les paramètres définissent le comportement des individus. Nous utilisons plus spécifiquement des processus de naissance et mort multi-types avec reproduction sexuée, compétition, et migration. Cette classe de modèles, leurs limites d'échelle et leurs applications à diverses questions d'écologie ont fait l'objet d'une littérature très riche, qui a démarré par la thèse de Nicolas Champagnat (Champagnat (2004)), et les articles Champagnat et Méléard (2007) et Champagnat *et al.* (2006) qui en sont issus.

Dans l'ensemble de nos articles Coron *et al.* (2018b, 2021, 2022), nous considérons une population d'individus haploïdes (i.e. qui portent une seule version de chaque gène), qui se reproduisent de façon sexuée mais hermaphrodite (i.e. sans distinction de sexes mâle/femelle), comme c'était déjà le cas dans le chapitre précédent. Ils sont caractérisés par leur génome et leur position spatiale, c'est-à-dire que ces deux éléments définissent leur comportement, qui consiste, à différents instants, à se déplacer sur un espace discret, se reproduire, ou mourir. Le génome d'un individu est transmis, de façon aléatoire et avec d'éventuelles mutations, lors de la reproduction. Cette population est modélisée par un processus de naissance et mort avec migration dont je détaille

ci-dessous l'espace d'états et les taux.

Espace d'états et changement d'échelle Chaque individu est caractérisé par un génome $g \in \mathcal{G}$ et une position (ou un patch) $i \in \mathcal{I}$. Les espaces $\mathcal{G}$ et $\mathcal{I}$ sont finis et seront définis ultérieurement car différents d'un travail à l'autre. Le type d'un individu est donc un élément de $\mathcal{E} = \mathcal{G} \times \mathcal{I}$ et la population est donc caractérisée à tout instant t par un vecteur $\mathbf{N}_t = (N_k(t))_{k \in \mathcal{G} \times \mathcal{I}} \in \mathbb{N}^{\mathcal{E}}$ qui donne les nombres respectifs d'individus de chaque type dans la population au temps t.

Ce processus sera considéré sous une hypothèse de grande taille de population. Cela signifie que l'on supposera que le nombre initial d'individus est de l'ordre de K, où K est un paramètre d'échelle voué à tendre vers l'infini et dont les différents paramètres du modèle dépendront. Le paramètre K a aussi une interprétation biologique, qui est la capacité de charge, c'est-à-dire l'ordre de grandeur du nombre d'individus que l'espace considéré peut supporter. Pour noter la dépendance en K de notre modèle, l'état de la population au temps t pourra être noté $\mathbf{N}_t^K$.

Migration On note $\rho(g, i, j, \mathbf{n})$ le taux auquel un individu quelconque de génotype g migre de la position $i \in \mathcal{I}$ à la position $j \in \mathcal{I}$ lorsque l'ensemble de la population est dans l'état $\mathbf{n} \in \mathbb{N}^{\mathcal{E}}$. La fonction ρ sera supposée Lipschitz en $\mathbf{n}$.

Reproduction et transmission génétique Chaque individu se reproduit à un taux β_g qui dépend de son génotype g, choisit un partenaire uniformément au hasard parmi les individus qui ont la même position que lui (ou vivent dans le même patch), mais cette rencontre donne lieu à la naissance (instantanée) d'un individu, avec une probabilité qui dépend du génotype des deux individus. Cette probabilité peut être symétrique ou non entre les génomes du premier parent (celui qui choisit de se reproduire) et du second parent (celui qui est choisi). Le génome de deux parents est transmis à leur enfant selon les lois de Mendel, en supposant que les loci sont indépendants, si besoin (Section 2.3.1 seulement). Lorsque la population est dans l'état $\mathbf{n} = (n_{g,i})_{g \in \mathcal{G}, i \in \mathcal{I}} \in \mathbb{N}^{\mathcal{E}}$, le taux auquel un individu de génotype g apparaît à la position i est donc de la forme :

$$b(g, i, \mathbf{n}) = \sum_{g_1, g_2 \in \mathcal{G}} \beta_{g_1} n_{g_1,i} \frac{n_{g_2,i}}{n_i} p_{g_1,g_2 -> g}$$

où $p_{g_1,g_2->g}$ est la probabilité pour qu'un couple d'individus ayant pour génotypes g_1, g_2 donne naissance à un individu de génotype g, et $n_i = \sum_{g \in \mathcal{G}} n_{g,i}$. Notons que cette probabilité inclut à la fois la transmission Mendélienne du génome (symétriques entre les deux génomes parentaux) et les éventuelles préférences d'appariement ou incompatibilités génétiques (non symétriques entre les deux génomes parentaux). En particulier cette probabilité n'est a priori pas symétrique en g_1 et g_2.

Mort Les individus de la population peuvent mourir soit de façon naturelle, soit du fait de la compétition avec les autres individus présents dans la population. On suppose que le génotype des individus n'influence pas ce comportement. On note $d(g, i, \mathbf{n}) = (d + \frac{c}{K} n_i) n_{g,i}$ le taux auquel

un individu quelconque ayant pour génotype g et position i meurt, lorsque l'ensemble de la population est dans l'état $\mathbf{n} = (n_{g,i})_{g\in\mathcal{G},i\in\mathcal{I}} \in \mathbb{N}^{\mathcal{E}}$. Le paramètre $d \in \mathbb{R}_+$ représente le taux de mort naturelle d'un individu, tandis que le paramètre $c/K \in \mathbb{R}_+^*$ représente le taux de compétition entre deux individus donnés. Rappelons que le nombre d'individus sera de l'ordre de K. Donc plus les individus sont nombreux, plus la compétition exercée par un individu donné sur un autre est faible ; en revanche la compétition totale exercée par l'ensemble des individus sur un individu donné reste toujours du même ordre.

Notons que les paramètres du modèle ne dépendent jamais de la position de l'individu considéré : on peut voir l'espace comme étant constitué de différents patchs qui sont dits écologiquement équivalents. Notre objectif essentiel est d'étudier les comportements qui émergent des préférences d'appariement contenues dans les paramètres $p_{g_1,g_2\to g}$, et éventuellement des paramètres de migration.

Changement d'échelle et convergence Notons $(\mathbf{e}_{g,i}, (g,i) \in \mathcal{G}\times\mathcal{I})$ la base canonique de $\mathbb{R}^{\mathcal{E}}$. La dynamique de la population considérée est représentée par la trajectoire d'un processus stochastique à valeurs dans $\mathbb{N}^{\mathcal{E}}$:

$$(\mathbf{N}^K(t), t \geq 0) = (N^K_{g,i}(t), (g,i) \in \mathcal{E}, t \geq 0),$$

dont les transitions sont données, pour tout $\mathbf{n} \in \mathbb{N}^{\mathcal{E}}$ et $(g,i) \in \mathcal{E}$, par :

$$\begin{array}{rcll} \mathbf{n} & \longrightarrow & \mathbf{n} + \mathbf{e}_{g,i} & \text{au taux } b(g,i,\mathbf{n}), \\ & \longrightarrow & \mathbf{n} - \mathbf{e}_{g,i} & \text{au taux } d(g,i,\mathbf{n}), \\ & \longrightarrow & \mathbf{n} + \mathbf{e}_{g,j} - \mathbf{e}_{g,i} & \text{au taux } \rho(g,i,j,\mathbf{n}). \end{array}$$

Comme mentionné précédemment nous allons considérer ce processus (qui sera défini plus spécifiquement dans chaque sous-section 2.3.1, 2.3.2 et 2.3.3) sous une échelle de grande taille de population. Plus précisément dans la suite nous supposerons que les tailles initiales de populations $(N^K_{\alpha,i}(0), (\alpha,i) \in \mathcal{E})$ sont d'ordre K. Nous allons donc naturellement nous intéresser au comportement du processus stochastique renormalisé :

$$Z^K = \frac{\mathbf{N}^K}{K}.$$

Notons $(\mathbf{z}^{(\mathbf{z^0})}(t), t \geq 0) = (z^{(\mathbf{z^0})}_{\alpha,i}(t), (\alpha,i) \in \mathcal{E})_{t\geq 0}$ l'unique solution de l'équation aux dérivées ordinaires

$$\frac{dz_{g,i}(t)}{dt} = b(g,i,\mathbf{z}(t)) - \Big(d + c\sum_{g\in\mathcal{G}} z_{g,i}(t)\Big) z_{g,i}(t) + \sum_{i'\in\mathcal{I}} (\rho(g,i',i,\mathbf{z}) - \rho(g,i,i',\mathbf{z})) \tag{2.1}$$

qui part de $\mathbf{z}^{(\mathbf{z}^0)}(0) = \mathbf{z^0} \in \mathbb{R}_+^{\mathcal{E}}$. L'unicité provient du fait que le champ de vecteur est localement Lipschitz et que les solutions n'explosent pas en temps fini (Chicone 2006). Alors le premier résultat obtenu (analogue au Théorème 1.3 du premier chapitre), découle de Ethier et Kurtz (1986) (Théorème 2.1 p. 456) :

Lemme 2.1. *Soit $T \in \mathbb{R}_+^*$. Supposons que la suite $(\mathbf{Z}^K(0), K \geq 1)$ converge en probabilité lorsque K tend vers l'infini vers un vecteur déterministe $\mathbf{z^0} \in (\mathbb{R}_+)^{\mathcal{E}}$. Alors*

$$\lim_{K\to\infty} \sup_{s\leq T} \|\mathbf{Z}^K(s) - \mathbf{z}^{(\mathbf{z^0})}(s)\| = 0 \quad \textit{en probabilité}, \tag{2.2}$$

où $\|.\|$ est la norme L^∞ sur $\mathbb{R}^{\mathcal{E}}$.

Ce résultat permet, lorsque K est grand, de déterminer le comportement du processus stochastique $\mathbf{N}^K$ à partir de celui du système dynamique (2.1), qui est plus facile à étudier et fait dans les articles Coron *et al.* (2018b) et Coron *et al.* (2021) l'objet d'une partie importante de notre travail.

2.3 Résultats

2.3.1 Émergence de l'homogamie

Motivation Notre premier but a été d'étudier les conditions d'émergence de l'homogamie. Ce travail, en collaboration avec Fabien Laroche (INRAE), est détaillé dans Coron *et al.* (2021). L'homogamie est le fait, pour un individu, de préférer se reproduire avec des individus qui lui ressemblent. Elle peut aussi prendre la forme d'une meilleure compatibilité génétique entre individus ayant des génomes proches, ou une meilleure fertilité de ces couples. Ces différentes formulations peuvent néanmoins se traduire par des modèles mathématiques différents donc il est important d'être vigilant sur les mots employés. Dans tous les cas on remarque que l'homogamie représente un avantage lorsqu'un individu est entouré d'individus comme lui, mais un coût lorsque son génotype (ou son phénotype) est peu représenté dans la population. L'homogamie est très répandue dans le monde vivant (McLain et Boromisa (1987); Herrero (2003); Savolainen *et al.* (2006)) et semble être un moteur majeur de spéciation (Gregorius (1992)).

Modèle Notre approche a consisté à définir un modèle d'homogamie le plus simple possible et à étudier les conditions nécessaires à l'installation à long terme de cette préférence d'appariement. Dans cette section il n'y a pas de structuration de l'espace : tous les individus vivent dans le même patch. On suppose que le comportement des individus est caractérisé par leur génome à deux loci bi-alléliques indépendants (i.e. qui sont situés sur deux chromosomes différents, ou suffisamment loin l'un de l'autre sur le génome, pour que leur transmission lors de la méiose se fasse de façon indépendante) : un locus qui code pour un phénotype et qui présente deux allèles : a et A, et un locus qui code pour une préférence d'appariement, portant sur ce premier locus, et qui présente aussi deux allèles : p et P. L'espace des génomes est ainsi $\mathcal{G} = \{AP, Ap, aP, ap\}$. Un individu qui porte l'allèle p est supposé ne pas avoir de préférence d'appariement, alors qu'un individu qui porte l'allèle P préfère se reproduire avec un individu qui porte le même allèle que lui (a ou A) au premier locus. Plus spécifiquement, on supposera que chaque couple d'individus se reproduit au même taux mais que la rencontre entre deux individus produit effectivement un

Parent 1	Parent 2	Taux de reproduction par couple
p	$---$	b_0/n
AP	A	b_+/n
aP	a	b_-/n

TABLE 2.1 – Taux de reproduction par couple en fonction du génotype de chaque parent, lorsque la taille de population est égale à n. L'homogamie exercée par le parent 1 porteur de l'allèle P se traduit mathématiquement par le fait que $b_- < b_0 < b_+$.

descendant avec une probabilité qui dépend de l'identité du premier et du second parent (qui n'exerce pas de préférence). Cela peut se traduire par le tableau de taux de reproduction 2.1, dans lequel on pourra noter le caractère dissymétrique des deux parents.

L'homogamie exercée par l'allèle P lorsqu'il est porté par le parent 1 se traduit mathématiquement par le fait que $b_- < b_0 < b_+$. On pose pour simplifier la suite : $b_0 = b$, $b_+ = b(1+\beta_1)$ et $b_- = b(1-\beta_2)$, où l'on suppose que $b > 0$, $\beta_1 \geq 0$ et $0 \leq \beta_2 \leq 1$. Enfin les génomes des deux parents sont transmis de façon Mendélienne, et en supposant que les deux loci considérés sont indépendants. Plus mathématiquement, cela veut dire qu'indépendamment pour chaque locus, l'allèle d'un des deux parents est choisi uniformément et est transmis à l'unique descendant produit. L'ensemble de ces hypothèses implique que le taux auquel un individu de génotype $i \in \mathcal{G}$ apparaît dans la population qui est dans l'état $\mathbf{n} = (n_{AP}, n_{Ap}, n_{aP}, n_{ap}) \in \mathbb{N}^4$ est donné par

$$\begin{aligned}
b_{AP}(\mathbf{n}) &= b\left[n_{AP} + \frac{1}{n}\left(\beta_1 n_{AP}\left(n_{AP} + \frac{n_{Ap}}{2}\right) - \beta_2\left(n_{AP}\left(n_{aP} + \frac{n_{ap}}{4}\right) + n_{Ap}\frac{n_{aP}}{4}\right)\right) + \frac{\Delta_{aP}}{2n}\right] \\
b_{Ap}(\mathbf{n}) &= b\left[n_{Ap} + \frac{1}{n}\left(\beta_1 n_{Ap}\frac{n_{AP}}{2} - \beta_2\left(n_{Ap}\frac{n_{aP}}{4} + n_{AP}\frac{n_{ap}}{4}\right)\right) - \frac{\Delta_{aP}}{2n}\right] \\
b_{aP}(\mathbf{n}) &= b\left[n_{aP} + \frac{1}{n}\left(\beta_1 n_{aP}\left(n_{aP} + \frac{n_{ap}}{2}\right) - \beta_2\left(n_{aP}\left(n_{AP} + \frac{n_{Ap}}{4}\right) + n_{ap}\frac{n_{AP}}{4}\right)\right) - \frac{\Delta_{aP}}{2n}\right] \\
b_{ap}(\mathbf{n}) &= b\left[n_{ap} + \frac{1}{n}\left(\beta_1 n_{ap}\frac{n_{aP}}{2} - \beta_2\left(n_{ap}\frac{n_{AP}}{4} + n_{aP}\frac{n_{Ap}}{4}\right)\right) + \frac{\Delta_{aP}}{2n}\right],
\end{aligned} \tag{2.3}$$

où

$$\Delta_{aP} := n_{aP}n_{Ap} - n_{AP}n_{ap}.$$

On suppose qu'avant le temps 0 la population n'était constituée que d'individus ayant pour allèle p au second locus, et donc se reproduisant uniformément, au taux b. Dans ce cas la taille de population est proche de son équilibre $(b-d)K/c$, où $b > d$ nécessairement (sinon la population résidente n'est pas viable). Au temps 0 apparaît un mutant, de génotype αP, avec $\alpha \in \{A, a\}$. On notera $\bar{\alpha}$ le complémentaire de α dans $\{A, a\}$ et on s'intéresse à l'émergence de l'allèle P dans ce contexte.

La dynamique de cette population est modélisée par un processus de naissance et mort multi-type avec compétition et reproduction sexuée, noté

$$(N^K(t), t \geq 0) := (N^K_{AP}(t), N^K_{Ap}(t), N^K_{aP}(t), N^K_{ap}(t), t \geq 0)$$

et qui prend ses valeurs dans $\mathbb{N}^4$. Comme mentionné précédemment, nous considérons plutôt le processus changé d'échelle :

$$(\mathbf{Z}^K(t), t \geq 0) := \left(\frac{N^K_{AP}(t)}{K}, \frac{N^K_{Ap}(t)}{K}, \frac{N^K_{aP}(t)}{K}, \frac{N^K_{ap}(t)}{K}, t \geq 0 \right),$$

de façon à nous placer sous une hypothèse de grande taille de population. Nous supposerons pour finir que $Z^K_{Ap}(0)/(Z^K_{Ap}(0) + Z^K_{ap}(0)) \to \rho_A$ en probabilité, ou encore que

$$\left(Z^K_{Ap}(0), Z^K_{ap}(0)\right) \underset{K\to\infty}{\to} \left(\rho_A \frac{b-d}{c}, (1-\rho_A)\frac{b-d}{c} \right)$$

en probabilité, comme mentionné précédemment.

Résultats Nous nous intéressons à l'invasion progressive de l'allèle homogame P qui vient d'apparaître, accompagnée de la disparition progressive de l'allèle p. Nos résultats portent sur trois éléments.

(*i*) Nous donnons une condition nécessaire et suffisante sur les paramètres de la dynamique de population et la composition initiale de la population pour que l'allèle P ait une probabilité strictement positive d'envahir la population (Proposition 2.2).

(*ii*) Nous caractérisons la probabilité d'invasion de l'allèle P, donnons une approximation de la durée de cette invasion et l'état final auquel elle conduit la population (Théorème 2.3).

(*iii*) Dans un cas particulier nous donnons la probabilité d'invasion du mutant homogame (Proposition 2.4).

Proposition 2.2 (CNS pour invasion avec probabilité > 0)**.** *L'allèle P envahit la population avec probabilité strictement positive si et seulement si :*

$$\beta_1 > \beta_2 \quad ou \quad \rho_A(1-\rho_A) < \frac{\beta_1(\beta_2+2)}{2(\beta_1+\beta_2)(\beta_1+2)}. \tag{2.4}$$

La Condition (2.4) donne deux conditions suffisantes (l'une d'entre elle au moins devant être réalisée) pour que la probabilité d'invasion du mutant soit strictement positive. La première impose que l'avantage conféré à la reproduction homogame soit plus important que le désavantage conféré à la reproduction hétérogame. La deuxième condition est qu'il y ait suffisamment peu de diversité allélique au premier locus, qui porte les allèles A et a. En particulier, même si l'avantage lié à l'homogamie est faible, si la proportion ρ_A est proche de 0 ou de 1 alors le mutant aura une probabilité strictement positive d'envahir la population.

La Condition (2.4) de la Proposition 2.2 découle du fait que le couple formé par les nombres $N_{AP}(t)$ et $N_{aP}(t)$ est, au début de sa dynamique, approximé par un processus de branchement dont la matrice moyenne est

$$J := \begin{pmatrix} \bar{\beta}_{AA} - b & \bar{\beta}_{Aa} \\ \bar{\beta}_{aA} & \bar{\beta}_{aa} - b \end{pmatrix} \tag{2.5}$$

où

$$\bar{\beta}_{\alpha\alpha} := \frac{b}{2}\left(1 + (\beta_1 + 1)\rho_\alpha - \frac{\beta_2}{2}\rho_{\bar{\alpha}}\right), \quad \bar{\beta}_{\alpha\bar{\alpha}} := \frac{b}{2}\left(1 - \frac{\beta_2}{2}\right)\rho_{\bar{\alpha}}. \tag{2.6}$$

Ce processus de branchement a une probabilité strictement positive de ne pas toucher 0 si et seulement si la matrice J a une valeur propre strictement positive, ce qui amène à la condition (2.4). Notons maintenant λ la valeur propre maximale de la matrice J qui est strictement positive sous la condition (2.4). Le résultat suivant donne, sous cette condition, la durée d'invasion de l'allèle P et l'état final de la population après cette invasion. Notons T_0^P le temps d'extinction de l'allèle P et T_{S_μ} le temps d'atteinte de $[\frac{b(1+\beta_1)-d}{c} - \mu, \frac{b(1+\beta_1)-d}{c} + \mu] \times \{0\} \times \{0\} \times \{0\}$ par la population.

Théorème 2.3. *Supposons que $\lambda > 0$, que $\rho_A \in (1/2, 1)$, et que pour un $\alpha \in \{A, a\}$*

$$\left(N_{\alpha P}^K(0), N_{\bar{\alpha} P}^K(0)\right) = (1, 0).$$

Alors il existe $q_\alpha \in [0, 1)$ solution d'une équation explicite, et une variable aléatoire B qui suit une loi de Bernoulli de paramètre $1 - q_\alpha$ telle que pour tout $0 < \mu < (b(1 + \beta_1) - d)/c$:

$$\lim_{K\to\infty} \left(\frac{T_{S_\mu} \wedge T_0^P}{\ln K}, \mathbf{1}_{\{T_{S_\mu} < T_0^P\}} \right) = B \times \left(\frac{1}{\lambda} + \frac{2}{b\beta_1}, 1 \right), \tag{2.7}$$

où la convergence a lieu en probabilité.
De plus,

$$\mathbf{1}_{\{T_0^P < T_{S_\mu}\}} \left\| \frac{\mathbf{N}^K(T_0^P)}{K} - (0, \rho_A, 0, 1 - \rho_A)\frac{b-d}{c} \right\|_1 \underset{K\to\infty}{\longrightarrow} 0 \qquad \text{en probabilité,} \tag{2.8}$$

où $\|\cdot\|_1$ est la norme L^1.

La preuve de ce théorème consiste à découper la dynamique du processus stochastique $(\mathbf{Z}^K(t), t \geq 0)$ durant l'invasion de l'allèle P en 3 phases : une phase mentionnée précédemment durant laquelle le processus $(N_{AP}(t), N_{aP}(t))_{t\geq 0}$ est approximé par un processus de branchement et le reste de la population est contrôlé, une phase durant laquelle le processus $(\mathbf{Z}^K(t), t \geq 0)$ est approximé par un système dynamique, et une phase d'extinction de l'allèle résident p et du génotype aP. Plus de détails sont donnés dans la Section 2.4. La variable aléatoire B de ce théorème est l'indicatrice de la survie d'un processus de branchement couplé au processus $(N_{AP}(t), N_{aP}(t))_{t\geq 0}$ qui donne la dynamique de la population mutante. Dans le cas où la condition (2.4) n'est pas vérifiée, la probabilité d'extinction q_α du mutant vaut 1, et la convergence

énoncée dans l'Équation (2.7) a lieu presque sûrement (vers $(0,0)$). Nous n'avons pas pu, en général, donner une formule explicite pour la probabilité d'extinction q_α de la population homogame, qui est définie comme solution d'une équation (plus de détails sont donnés dans la Section 2.4 qui rassemble des éléments de preuves pour ce chapitre, cf l'Équation (2.26) pour ce point précis). Cependant dans le cas particulier où il n'y a que des individus A ou a dans la population avant l'arrivée de l'allèle P, nous pouvons donner sa valeur.

Proposition 2.4. *Supposons que $\rho_A = 1$, c'est-à-dire qu'il n'y a que des individus Ap avant l'arrivée de l'allèle P dans la population. Dans ce cas*

$$q_A = \frac{2}{2+\beta_1}$$

et

$$q_a = \frac{1}{2-\beta_2}\left(\frac{6-\beta_1\beta_2+4\beta_1-\beta_2}{2+\beta_1} - \sqrt{\left(\frac{6-\beta_1\beta_2+4\beta_1-\beta_2}{2+\beta_1}\right)^2 - 4(2-\beta_2)}\right).$$

Ce résultat est complété par des simulations numériques (Figure 2.1) qui montrent dans le cas général où ρ_A est quelconque, une dépendance complexe des probabilités d'extinction (q_A, q_a) en les paramètres ρ_A, β_1 et β_2, et en particulier une non différentiabilité de q_A en fonction de la proportion initiale d'allèle A, ρ_A, pour certaines valeurs de β_1 et β_2 (autour de la criticalité, c'est-à-dire lorsque q_1 passe de 1 à une valeur strictement plus petite).

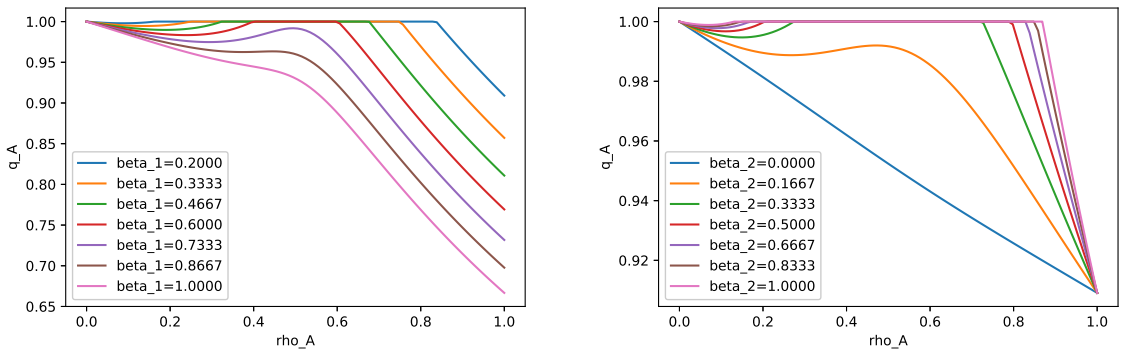


FIGURE 2.1 – Valeurs de q_A en fonction de ρ_A pour différentes valeurs de β_1 et β_2. À gauche, β_2 est fixé à 0.7 et β_1 varie. À droite, β_1 est fixé à 0.2 et β_2 varie. Dans les deux cas $b = 1$. La symétrie de notre modèle implique que $q_a(\rho_A) = q_A(1-\rho_A)$.

2.3.2 Homogamie, recherche de partenaire, et spéciation

Motivation et modèle Notre deuxième objectif a été d'étudier comment l'homogamie, couplée à une structuration spatiale, peut générer l'arrêt des flux de gènes, ou des reproductions,

entre deux sous-populations. Pour ce travail, détaillé dans l'article Coron *et al.* (2018b), nous supposons maintenant que les individus sont tous homogames, que l'espace des génomes est constitué de deux allèles possibles, correspondant à deux versions d'un gène, situé à un locus donné du génome : $\mathcal{G} = \{A, a\}$, et que l'espace est constitué de deux patchs (écologiquement équivalents, comme mentionné dans la Section 2.2) : $i \in \mathcal{I} = \{1, 2\}$. On suppose alors que

(i) Les individus se rencontrent pour se reproduire, de façon uniforme au sein de chaque patch, mais la probabilité pour qu'une rencontre entre deux individus donne lieu à un descendant est plus élevée lorsqu'ils ont le même génotype. Ainsi le taux auquel un individu de génotype α naît dans le patch i vaut

$$\begin{aligned}\lambda_{\alpha,i}(\mathbf{n}) =& b\left(n_{\alpha,i}\beta\frac{n_{\alpha,i}}{n_{\alpha,i}+n_{\bar{\alpha},i}} + \frac{1}{2}n_{\alpha,i}\frac{n_{\bar{\alpha},i}}{n_{\alpha,i}+n_{\bar{\alpha},i}} + \frac{1}{2}n_{\bar{\alpha},i}\frac{n_{\alpha,i}}{n_{\alpha,i}+n_{\bar{\alpha},i}}\right)\\ =& bn_{\alpha,i}\frac{\beta n_{\alpha,i}+n_{\bar{\alpha},i}}{n_{\alpha,i}+n_{\bar{\alpha},i}}.\end{aligned} \tag{2.9}$$

Le paramètre $\beta > 1$ représente la compatibilité génétique d'un couple d'individus homogame, dans le sens où les individus se rencontrent uniformément au hasard et deux individus qui se rencontrent ont une probabilité β fois plus élevée de produire un descendant s'ils ont le même génome.

(ii) Les individus migrent d'un patch à l'autre à un taux proportionnel à la proportion d'individus qui ne sont pas du même génotype qu'eux, dans leur patch (cf Figure 2.2) :

$$\rho_{\alpha,i\to\bar{i}}(\mathbf{n}) = pn_{\alpha,i}\left(1 - \frac{n_{\alpha,i}}{n_{\alpha,i}+n_{\bar{\alpha},i}}\right) = p\frac{n_{\alpha,i}n_{\bar{\alpha},i}}{n_{\alpha,i}+n_{\bar{\alpha},i}}. \tag{2.10}$$

Les exemples d'espèces d'animaux qui migrent pour trouver des partenaires sexuels sont bien documentés (Schwagmeyer 1988; Höner *et al.* 2007). Par ailleurs un mécanisme similaire de migration a été étudié dans Payne et Krakauer (1997) pour un espace continu. Enfin nous avons considéré dans l'article Coron *et al.* (2018b) des classes plus générales de modèle, avec plus de deux patchs et des formes plus générales de migration.

Résultats Notre premier résultat est le Lemme 2.1, énoncé en préambule de ce chapitre. Il nous dit que lorsque la capacité de charge K tend vers l'infini, le processus stochastique

$$(\mathbf{Z}^K(t), t \geq 0) = (Z^K_{\alpha,i}(t), (\alpha, i) \in \mathcal{E}, t \geq 0) = \left(\frac{\mathbf{N}^K(t)}{K}, t \geq 0\right),$$

est proche de l'unique solution $(\mathbf{z}(t), t \geq 0)$ du système dynamique suivant, qui part de la condition initiale $\mathbf{z}(\mathbf{0}) := \lim_{\mathbf{K}\to\infty}(\mathbf{N}^{\mathbf{K}}_{\alpha,\mathbf{i}}(\mathbf{0})/\mathbf{K})_{(\alpha,\mathbf{i})\in\mathcal{E}} \in \mathbb{R}^{\mathcal{E}}_+$:

$$\left\{\begin{array}{l}
\frac{d}{dt}z_{A,1}(t) = z_{A,1}\left[b\frac{\beta z_{A,1}+z_{a,1}}{z_{A,1}+z_{a,1}} - d - c(z_{A,1}+z_{a,1}) - p\frac{z_{a,1}}{z_{A,1}+z_{a,1}}\right] + p\frac{z_{A,2}z_{a,2}}{z_{A,2}+z_{a,2}}\\
\frac{d}{dt}z_{a,1}(t) = z_{a,1}\left[b\frac{\beta z_{a,1}+z_{A,1}}{z_{A,1}+z_{a,1}} - d - c(z_{A,1}+z_{a,1}) - p\frac{z_{A,1}}{z_{A,1}+z_{a,1}}\right] + p\frac{z_{A,2}z_{a,2}}{z_{A,2}+z_{a,2}}\\
\frac{d}{dt}z_{A,2}(t) = z_{A,2}\left[b\frac{\beta z_{A,2}+z_{a,2}}{z_{A,2}+z_{a,2}} - d - c(z_{A,2}+z_{a,2}) - p\frac{z_{a,2}}{z_{A,2}+z_{a,2}}\right] + p\frac{z_{A,1}z_{a,1}}{z_{A,1}+z_{a,1}}\\
\frac{d}{dt}z_{a,2}(t) = z_{a,2}\left[b\frac{\beta z_{a,2}+z_{A,2}}{z_{A,2}+z_{a,2}} - d - c(z_{A,2}+z_{a,2}) - p\frac{z_{A,2}}{z_{A,2}+z_{a,2}}\right] + p\frac{z_{A,1}z_{a,1}}{z_{A,1}+z_{a,1}}.
\end{array}\right. \tag{2.11}$$

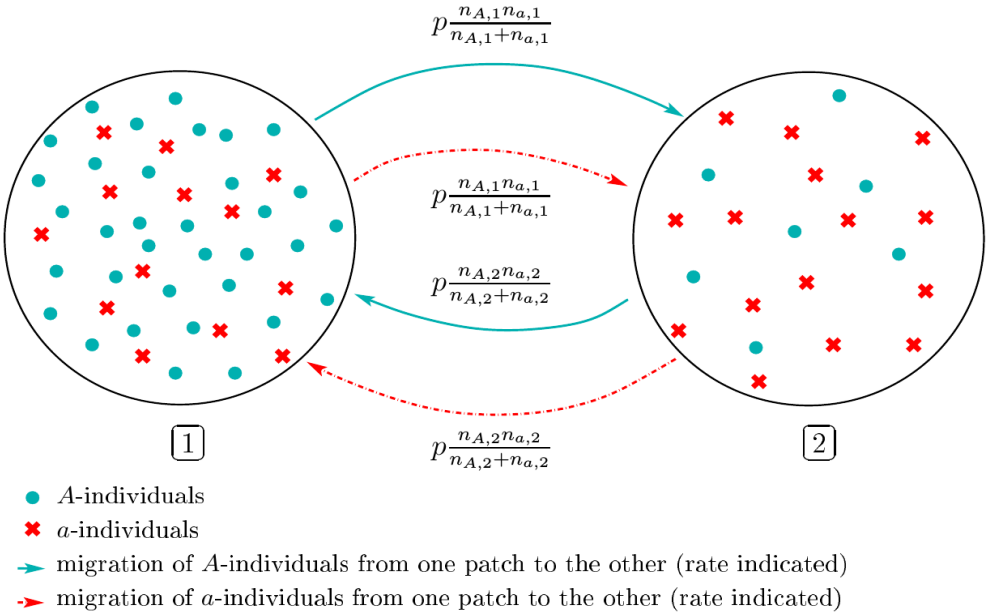


FIGURE 2.2 – Taux auquel une migration entre deux patchs a lieu.

Notre deuxième résultat consiste à étudier les équilibres du système dynamique (2.11). Rappelons que $\beta > 1$, $b > d$, et notons

$$\zeta := \frac{\beta b - d}{c}. \tag{2.12}$$

Théorème 2.5. *Les points d'équilibre dans $\mathbb{R}^{\mathcal{E}}_+$ du système dynamique (2.11) sont les suivants :*

1. *Les points pour lesquels la population ne contient plus qu'un seul type dans un seul patch*

$$(\zeta,0,0,0) \quad (0,\zeta,0,0) \quad (0,0,\zeta,0) \quad (0,0,0,\zeta) \tag{2.13}$$

2. *Les points pour lesquels chaque type est présent dans exactement un patch :*

$$(\zeta,0,0,\zeta), \quad (0,\zeta,\zeta,0) \tag{2.14}$$

3. *Les points pour lesquels seulement un type reste présent, dans les deux patchs :*

$$(\zeta,0,\zeta,0), \quad (0,\zeta,0,\zeta) \tag{2.15}$$

4. *Les points pour lesquels les deux types restent présents dans les deux patchs :*

$$\left(\frac{b(\beta+1)-2d}{4c}, \frac{b(\beta+1)-2d}{4c}, \frac{b(\beta+1)-2d}{4c}, \frac{b(\beta+1)-2d}{4c}\right) \tag{2.16}$$

$$\left(\frac{\zeta+\sqrt{\Delta}}{2}, \frac{\zeta-\sqrt{\Delta}}{2}, \widetilde{\zeta}, \widetilde{\zeta}\right), \quad \left(\frac{\zeta-\sqrt{\Delta}}{2}, \frac{\zeta+\sqrt{\Delta}}{2}, \widetilde{\zeta}, \widetilde{\zeta}\right), \tag{2.17}$$

$$\left(\widetilde{\zeta}, \widetilde{\zeta}, \frac{\zeta+\sqrt{\Delta}}{2}, \frac{\zeta-\sqrt{\Delta}}{2}\right), \quad \left(\widetilde{\zeta}, \widetilde{\zeta}, \frac{\zeta-\sqrt{\Delta}}{2}, \frac{\zeta+\sqrt{\Delta}}{2}\right). \tag{2.18}$$

Les seuls équilibres stables parmi cette liste sont ceux définis dans les Équations (2.14) *et* (2.15).

Notre troisième résultat porte alors sur la convergence des solutions du système dynamique (2.11) vers l'équilibre qui nous intéresse $(\zeta, 0, 0, \zeta)$, pour lequel chaque type finit dans un seul patch (les équilibres $(0, \zeta, \zeta, 0)$, $(\zeta, 0, \zeta, 0)$ et $(0, \zeta, 0, \zeta)$ pouvant être traités de façon symétrique). Pour cela nous introduisons le domaine

$$\mathcal{D} := \{\mathbf{z} \in \mathbb{R}_+^{\mathcal{E}}, z_{A,1} - z_{a,1} > 0, z_{a,2} - z_{A,2} > 0\},$$

et le nombre réel positif

$$p_0 = \frac{\sqrt{b(\beta-1)[b(3\beta+1) - 4d]} - b(\beta-1)}{2}. \tag{2.19}$$

Nous savons que $p_0 < b(\beta+1) - 2d$ sous les hypothèses appropriées. Enfin, pour tout $p \in [0, b(\beta+1) - 2d]$, nous introduisons le domaine

$$\mathcal{K}_p := \left\{ \mathbf{z} \in \mathcal{D},\ \{z_{A,1} + z_{a,1},\ z_{A,2} + z_{a,2}\} \in \left[\frac{b(\beta+1) - 2d - p}{2c}, \frac{2b\beta - 2d + p}{2c} \right] \right\}.$$

Théorème 2.6. *Soit $p < p_0$. On a*

(i) Toute solution $(\mathbf{z}(t), t \geq 0)$ du système dynamique(2.11) qui démarre dans $\mathcal{D}$ converge vers l'équilibre $(\zeta, 0, 0, \zeta)$.

(ii) Si la condition initiale $\mathbf{z}(0)$ de la solution $(\mathbf{z}(t), t \geq 0)$ du système dynamique (2.11) est dans $\mathcal{K}_p$, alors il existe deux constantes positives k_1 et k_2, dépendantes de la conditions initiales, telles que pour tout $t \geq 0$,

$$\|\mathbf{z}(t) - (\zeta, 0, 0, \zeta)\| \leq k_1 e^{-k_2 t}.$$

Enfin notre quatrième résultat porte sur le processus stochastique $\mathbf{Z}^K(t), t \geq 0)$: son comportement en temps long et le temps au bout duquel chaque type reste présent dans un patch uniquement.

Théorème 2.7. *Supposons que $\mathbf{Z}^K(0)$ converge en probabilité vers un vecteur déterministe $\mathbf{z^0}$ appartenant à $\mathcal{D}$, avec $(z^0_{a,1}, z^0_{A,2}) \neq (0, 0)$. Soit*

$$\mathcal{B}_\varepsilon := [(\zeta - \varepsilon)K, (\zeta + \varepsilon)K] \times \{0\} \times \{0\} \times [(\zeta - \varepsilon)K, (\zeta + \varepsilon)K].$$

Il existe trois constantes positives ε_0, C_0 et m, et une constante positive V dependant de (m, ε_0) telles que si $p < p_0$ et $\varepsilon \leq \varepsilon_0$, alors

$$\lim_{K\to\infty} \mathbb{P}\left(\left| \frac{T^K_{\mathcal{B}_\varepsilon}}{\log K} - \frac{1}{b(\beta-1)} \right| \leq C_0\varepsilon,\ \mathbf{N}^K\left(T^K_{\mathcal{B}_\varepsilon} + t\right) \in \mathcal{B}_{m\varepsilon}\ \forall t \leq e^{VK} \right) = 1, \tag{2.20}$$

où $T^K_{\mathcal{B}}$, $\mathcal{B} \subset \mathbb{R}_+^{\mathcal{E}}$ est le temps d'atteinte de l'ensemble $\mathcal{B}$ par le processus $\mathbf{N}^K$.

Ce théorème donne l'ordre de grandeur du temps nécessaire pour atteindre l'isolement reproductif entre les deux patchs, en fonction du paramètre d'échelle de la taille de la population, K. Notons que le temps nécessaire pour atteindre l'isolement reproductif ne dépend pas du paramètre p. Cela s'explique par le fait que le temps nécessaire pour atteindre un voisinage de l'état $(\zeta, 0, 0, \zeta)$ est de l'ordre de 1, et qu'à partir de ce voisinage, le temps nécessaire pour l'extinction complète des individus a dans le patch 1 et des individus A dans le patch 2 est beaucoup plus long (de l'ordre de $C \log K$). Au cours de cette seconde phase, les migrations entre les deux patchs sont déjà équilibrées, ce qui entraîne l'indépendance par rapport à p. De plus, la constante C ne dépend pas de d et c puisqu'il n'y a pas de différence écologique entre les deux types et les deux patchs : durant la seconde phase, le taux de natalité des individus a dans le patch 1 est proche de b puisque le patch 1 est presque entièrement occupé par des individus de génotype A, et leur taux de mortalité naturelle peut être approximé par $d + c\zeta = b\beta$ où le terme $c\zeta$ provient de la compétition exercée par les individus A. Ainsi, leur taux de croissance naturelle est approximativement $b - b\beta$ qui dépend bien uniquement des paramètres de naissance.

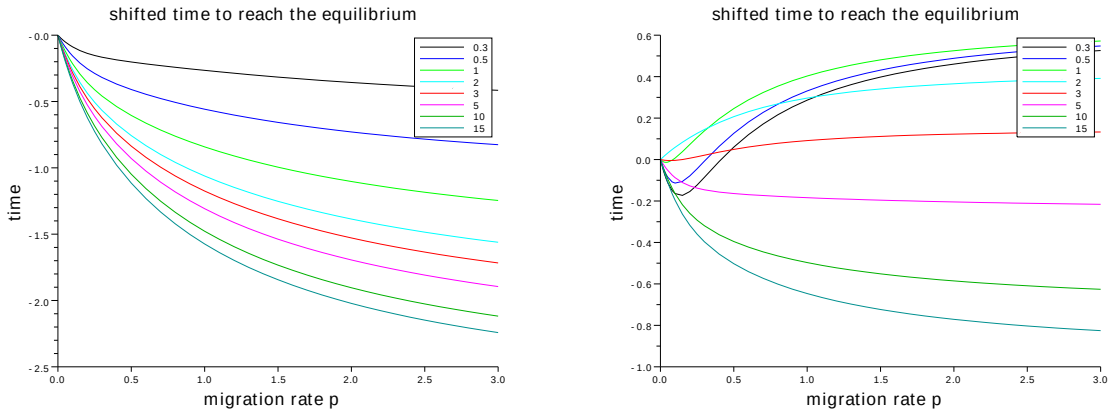
Le Théorème 2.7 donne non seulement une estimation du temps auquel la population atteint un voisinage de la limite du système dynamique donnée dans l'Équation (2.14), mais indique aussi qu'après ce temps la population reste dans ce voisinage pendant longtemps. Le Théorème 4 de Coron *et al.* (2018b) généralise le Théorème 2.7 au cas où il y a plus de 2 patchs. Enfin, l'hypothèse $(z^0_{a,1}, z^0_{A,2}) \neq (0, 0)$ est nécessaire pour obtenir la borne inférieure dans (2.20). En effet, si $(z^0_{a,1}, z^0_{A,2}) = (0, 0)$, l'ensemble $\mathcal{B}_\varepsilon$ est atteint plus rapidement, et donc seule la borne supérieure reste valable. Dans ce cas, la vitesse d'atteinte de l'ensemble $\mathcal{B}_\varepsilon$ dépendra de la vitesse de convergence de la suite $(Z^K_{a,1}, Z^K_{A,2})$ vers $(0, 0)$. Dans l'exemple trivial où $(Z^K_{a,1}, Z^K_{A,2}) = (0, 0)$, $T^K_{\mathcal{B}}$ sera d'ordre 1, ce qui correspond au temps nécessaire aux processus $Z^K_{A,1}$ et $Z^K_{a,2}$ pour atteindre chacun un voisinage de l'équilibre ζ.

Notons que la limite atteinte dépend du génotype initialement majoritaire dans chaque patch, puisque le sous-ensemble $\mathcal{D}$ est invariant sous le système dynamique (2.11). Par ailleurs, lorsque $p = 0$, les résultats du Théorème 2.6 peuvent être prouvés facilement puisque les deux patchs sont indépendants l'un de l'autre. La difficulté est donc de prouver le résultat lorsque $p > 0$. Notre argumentation nous permet de déduire une constante explicite p_0 sous laquelle on a une convergence vers un équilibre avec isolement reproductif entre patchs. Cependant, nous ne sommes pas en mesure de déduire un résultat rigoureux pour tout p. En effet, lorsque p augmente, il y a plus de mélanges entre les deux patchs ce qui rend le modèle difficile à étudier. Néanmoins, les simulations présentées dans la Figure 2.3 suggèrent que le résultat reste vrai.

Dans cette figure, nous traçons le temps $T_\varepsilon(p)$ auquel la solution du système dynamique (2.11) atteint l'ensemble

$$\mathcal{S}_\varepsilon = \left\{(z_{A,1}, z_{a,1}, z_{A,2}, z_{a,2}) \in \mathbb{R}^4_+, (z_{A,1} - \zeta)^2 + z^2_{a,1} + z^2_{A,2} + (z_{a,2} - \zeta)^2 \leq \varepsilon^2\right\},$$

c'est-à-dire le premier temps auquel cette solution atteint un $\varepsilon-$voisinage de $(\zeta, 0, 0, \zeta)$, pour différentes conditions initiales et en fonction du taux de migration p. Les paramètres démographiques sont $\beta = 2$, $b = 2$, $d = 1$ et $c = 0.1$, et nous prenons $\varepsilon = 0.01$. Pour ces paramètres,



(a) $(z_{A,2}(0), z_{a,2}(0)) = (1, 30)$ (b) $(z_{A,2}(0), z_{a,2}(0)) = (15, 16)$

FIGURE 2.3 – Pour différentes conditions initiales nous traçons $p \mapsto T_\varepsilon(p) - T_\varepsilon(0)$. La condition initiale est $(z_{A,1}(0), z_{A,1}(0) - 0.1, z_{A,2}(0), z_{a,2}(0))$ où $z_{A,1}(0) \in \{0.3, 0.5, 1, 2, 3, 5, 10, 15\}$ comme représenté par les couleurs de la légende, et $(z_{A,2}(0), z_{a,2}(0)) = (1, 30)$ sur la gauche, et $(z_{A,2}(0), z_{a,2}(0)) = (15, 16)$ sur la droite.

$\zeta = 30$ et $p_0 = \sqrt{5} - 1 \simeq 1.24$. Nous remarquons que l'impact du taux de migration sur le comportement du temps nécessaire à la population pour atteindre un voisinage de son équilibre correspondant à un isolement reproducteur est fortement dépendant de la condition initiale.

2.3.3 Hétérogamie et diversité génétique

Motivation et modèle La troisième question qui nous a intéressées est celle de la diversité génétique générée par les préférences d'appariement et en particulier par l'hétérogamie. Ce travail, en collaboration avec Violaine Llaurens (CNRS), est détaillé dans Coron *et al.* (2022). Des résultats classiques de génétique des populations montrent que la surdominance, c'est-à-dire l'avantage sélectif des individus hétérozygotes sur les individus homozygotes, favorise la diversité génétique (Lewontin *et al.* 1978). Cette surdominance peut être le résultat de préférences d'appariement, lorsque les individus préfèrent se reproduire avec des individus éloignés génétiquement (Maisonneuve *et al.* 2021). C'est le phénomène que nous étudions ici.

Pour cela, comme dans la Section 2.3.2, nous considérons une population d'individus haploïdes caractérisés par leur génome à un seul locus. Plus précisément nous supposons qu'il y a k allèles possibles au locus considéré, notés $1, 2, ..., k$. Ces individus se reproduisent de façon sexuée : chaque individu de type i se reproduit au taux β_i, choisit un partenaire uniformément dans la population (qui n'est plus spatialisée, comme dans la Section 2.3.1), et cette rencontre donne lieu à un nouvel individu avec une probabilité p_{ij} si le deuxième parent a le génome (ou le type) j. Nous supposons pour finir une transmission Mendélienne de ces allèles, comme dans les sections précédentes (i.e. l'enfant d'un couple de parents de génotypes respectifs i et j aura pour génotype

i ou j avec probabilité 1/2 pour chacun). On sait alors, sous une hypothèse de grande taille de population (Lemme 2.1 encore), que le processus de naissance et mort considéré et proprement renormalisé converge vers l'unique solution $(z_1(t), z_2(t), ..., z_k(t))_{t\geq 0}$ du système dynamique

$$\dot{z}_i(t) = z_i(t)\left(\sum_{j=1}^{k} \frac{\beta_i p_{ij} + \beta_j p_{ji}}{2} \frac{z_j(t)}{z(t)} - d - cz(t)\right), \quad i \in \{1, ..., k\},\ t \geq 0 \tag{2.21}$$

partant de $(z_1(0), ..., z_k(0)) \in \mathbb{R}_+^k$, où pour tout $t > 0$, $z(t) = \sum_{i=1}^k z_i(t)$ est la masse totale de population au temps t.

Nous notons

$$b := \inf_{1\leq i,j\leq k} \frac{\beta_i p_{ij} + \beta_j p_{ji}}{2}$$

et supposons que $b \geq d > 0$. Pour $(i, j) \in \{1, ..., k\}^2$, on introduit aussi

$$s_{ij} := \frac{\beta_i p_{ij} + \beta_j p_{ji}}{2b} - 1,$$

et l'Équation (2.21) se réécrit alors

$$\dot{z}_i(t) = z_i(t)\left(b\sum_{j=1}^{k}(1 + s_{ij})\frac{z_j(t)}{z(t)} - d - cz(t)\right). \tag{2.22}$$

Pour tout $i, j \in \{1, .., k\}$, le paramètre s_{ij} peut être interprété comme l'avantage sélectif d'une paire de parents ayant pour génotypes i et j. On note $M = (s_{ij})_{1\leq i,j\leq k}$ la matrice de l'ensemble des avantages sélectifs.

Résultat général Notre premier résultat donne une condition sur les avantages sélectifs s_{ij} sous laquelle la diversité génétique sera maintenue, ainsi que l'état limite de la population considérée, sous cette condition.

Théorème 2.8. *Supposons que* $\det(M) \neq 0$ *et que*

$$M^{-1}\mathbf{1} > 0, \qquad \text{où} \qquad \mathbf{1} = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}. \tag{2.23}$$

Le système dynamique (2.22) *admet un unique équilibre strictement positif*

$$Z^* := \frac{1}{c}\left(b + \frac{b}{\mathbf{1}^T M^{-1}\mathbf{1}} - d\right)\frac{M^{-1}\mathbf{1}}{\mathbf{1}^T M^{-1}\mathbf{1}}, \tag{2.24}$$

où $\mathbf{1}^T$ *est le vecteur* $\mathbf{1}$ *transposé.*
Qui plus est, en partant de n'importe quelle distribution allélique strictement positive, la population se stabilisera autour de cet équilibre si et seulement si la matrice M *a exactement* 1 *valeur propre strictement positive et* $k - 1$ *valeurs propres strictement négatives.*

La Condition (2.23) est facile à vérifier numériquement, la matrice M étant donnée. Le Théorème 2.8 permet en outre l'exploration de la diversité génétique (typiquement le nombre d'allèles maintenus en temps long) permise par différents structures de préférences d'appariement et notamment d'hétérogamie. C'est l'objet de la fin de cette section.

Application à quelques cas particuliers Nous commençons par l'application du Théorème 2.8 à quelques cas particuliers, permettant de retrouver ou de démontrer plusieurs résultats déjà connus ou conjecturés par la littérature de génétique des populations. Nous prouvons en effet que dans le cas où la population est constituée de deux allèles, alors la population admet un équilibre strictement positif si et seulement si $s_{12} > s_{11}$ et $s_{21} > s_{22}$, c'est-à-dire si tous les individus sont hétérogames. Ce résultat était connu (Kimura et Ohta (2020)). Dans le cas où la population est constituée de trois allèles nous montrons que ces trois allèles se maintiennent en temps long si et seulement si

$$s_{12} < s_{13} + s_{23}\,, \quad s_{13} < s_{12} + s_{23}\,, \quad s_{23} < s_{12} + s_{13}. \tag{2.25}$$

Dans l'article Lewontin *et al.* (1978) les auteurs ont prouvé que dans une population constituée de k allèles une condition circulaire du même type est nécessaire au maintien de la diversité génétique. Dans le cas $k = 3$ nous montrons donc que cette condition est aussi suffisante. Enfin le Théorème 2.8 permet de montrer que dans le cas le plus simple d'hétérogamie, tel que $s_{ij} = \rho + \epsilon \mathbf{1}_{i \neq j}$ alors la diversité génétique se maintient en temps long quel que soit le nombre d'allèles initialement présents.

Construction de la diversité génétique Dans cette section nous appliquons le Théorème 2.8 à l'étude de la construction progressive de la diversité génétique, par apparition successive de mutations. Nous supposons donc que de nouveaux allèles peuvent apparaître dans une population dans laquelle un ou plusieurs allèles coexistent déjà. Nous qualifions donc le nouvel allèle apparu de mutant et les allèles préexistants d'allèles résidents. Nous considérons que les mutations sont suffisamment rares pour que la dynamique de la population résidente atteigne son équilibre entre deux apparitions de mutations. Nous cherchons alors à étudier le devenir des mutations successives et non simultanées dans la population. Ce cadre classique est proche du cadre de la dynamique adaptative introduit dans Metz *et al.* (1996), car nous considérons des événements de mutation rares. Cependant, nous ne supposons pas que les mutations ont nécessairement de faibles effets. Nous caractérisons les conditions sur les paramètres d'avantage sélectif de l'allèle mutant (lorsqu'il s'accouple aux différents allèles résidents), qui permettent son invasion, c'est-à-dire sa persistance à long terme dans la population.

Théorème 2.9. *Considérons une population résidente stable qui contient k types et caractérisée par une matrice de préférences M, c'est-à-dire (d'après le Théorème 2.8) telle que M satisfait $M^{-1}\mathbf{1} > 0$ et la deuxième valeur propre de M est négative.*
Considérons un type mutant qui arrive dans la population résidente, caractérisé par les avantages

sélectifs $S = (s_{k+1,i})_{i=1,..,k}$ *et* $\sigma = s_{k+1,k+1}$. *Notons la nouvelle matrice de préférences*

$$\bar{M} = \begin{pmatrix} M & S \\ S^T & \sigma \end{pmatrix}.$$

Si $\bar{M}^{-1}\mathbf{1} > 0$, *c'est-à-dire si l'équilibre à* $k+1$ *types existe, alors il est globalement asymptotiquement stable.*

Ce résultat donne le comportement asymptotique de la population lorsqu'un équilibre à $k+1$ allèles existe. Lorsque cet équilibre n'existe pas l'apparition du dernier allèle peut donner lieu à des comportements divers, notamment la disparition d'un ou plusieurs allèles pré-existants. Ce phénomène est illustré dans la Figure 2.4 dans un cas très simple où il y a 3 allèles résidents et où la matrice M avant et après apparition du quatrième allèle vaut respectivement

$$\begin{pmatrix} 0 & s & s \\ s & 0 & s \\ s & s & 0 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 0 & s & s & x \\ s & 0 & s & y \\ s & s & 0 & z \\ x & y & z & 0 \end{pmatrix}.$$

La figure 2.4 montre qu'après l'apparition du quatrième allèle, lorsque l'équilibre à 4 allèles n'existe pas, l'état final de la population peut contenir 2 ou 3 allèles.

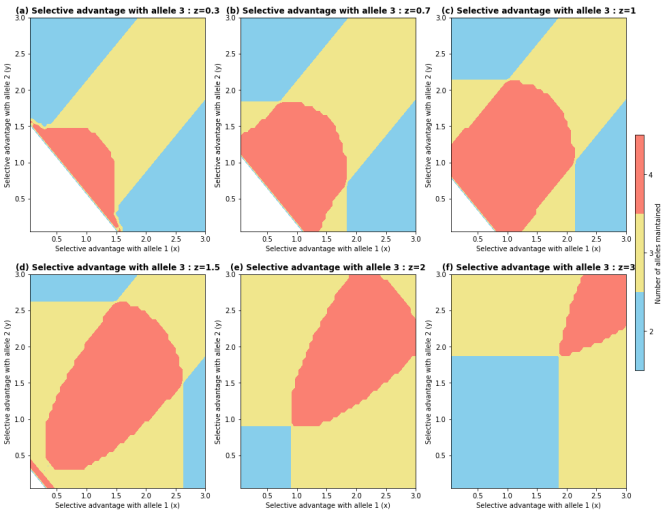


FIGURE 2.4 – Nombre d'allèles dans la population à l'équilibre, après l'introduction d'un nouvel allèle dans une population d'hétérogamie initialement homogène (paramètre s) en fonction des paramètres de préférences d'appariement entre l'allèle mutant et chacun des allèles résidents. Dans la zone blanche le mutant n'envahit pas, dans la zone rouge les quatre allèles se maintiennent, dans la zone jaune seulement 3 allèles se maintiennent, et dans la zone bleue seulement 2 allèles se maintiennent. Les paramètres sont $b = 1, d = 0, c = 1, s = 1$.

Application à un modèle de génétique quantitative Notre dernier résultat porte sur l'étude d'un modèle de génétique quantitative. Dans cette section l'espace des génomes est $\mathcal{G} = \{0,1\}^L$ et on définit la distance entre deux génomes $g = (g_1, g_2, ..., g_L)$ et $g' = (g'_1, g'_2, ..., g'_L) \in \mathcal{G}$ par

$$d(g, g') = \sum_{i=1}^{L} \mathbf{1}_{g_i \neq g'_i}.$$

Cette hypothèse de structure du génome en L sites polymorphes est appropriée par exemple pour modéliser le complexe majeur d'histocompatibilité (*MHC*) chez les vertébrés (Stefan *et al.* 2019). On suppose alors, pour rester dans le cadre d'étude de l'hétérogamie, que plus deux individus ont des génomes distants plus le succès reproductif du couple qu'ils forment est élevé. Plus précisément nous supposons que $s_{xy} = d(x,y)^\alpha$ et nous étudions l'impact de α et de la taille L du génome sur la diversité génétique soutenue par la population. Nous obtenons que le paramètre α joue un rôle très important dans la quantité de diversité génétique que la population peut contenir. Pour commencer, nous montrons que quel que soit le nombre de sites L, l'équilibre avec tous les génomes de $\mathcal{G}$ existe, mais il semble instable lorsque $\alpha \geq 1$ (Figure 2.5).

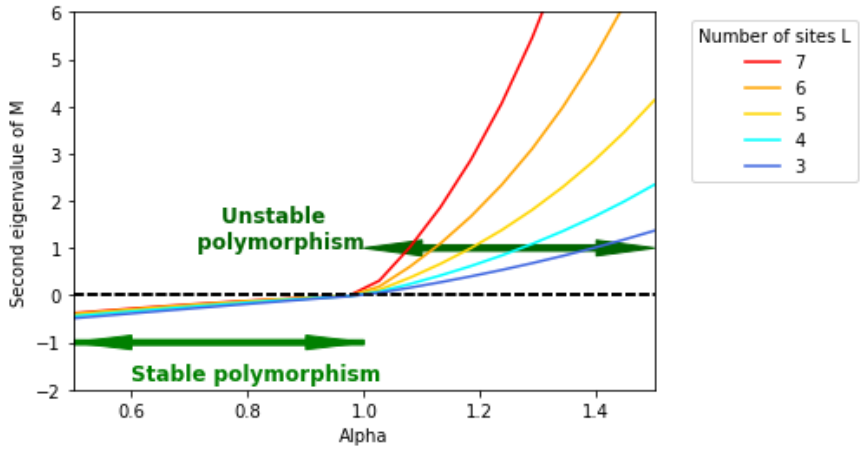
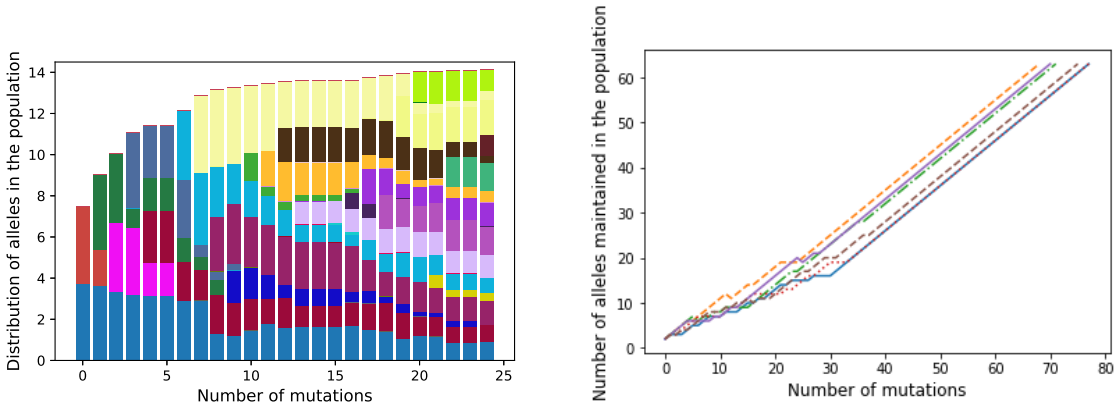


FIGURE 2.5 – **Stabilité de la population dans laquelle tous les allèles possibles sont présents**, explorée en utilisant la deuxième valeur propre de la matrice de sélection M, pour différents nombres de sites L au locus considéré, et en fonction du paramètre α de la fonction f qui code la relation entre la distance génétique et la préférence d'appariement. Notons que dès que $\alpha \geq 1$, cette valeur propre devient positive, ce qui implique que l'équilibre contenant tous les allèles n'est pas stable, d'après le Théorème 2.8.

Lorsque $\alpha < 1$, le Théorème 2.8 nous dit que l'équilibre avec tous les génomes est stable. Néanmoins ce résultat ne nous permet pas de savoir quel sera le résultat d'une introduction progressive de mutations. Cette question est difficile et nous l'avons étudiée de façon numérique. La Figure 2.6 donne un résultat de simulation de la dynamique du nombre d'allèles dans la population au cours du temps et au fur et à mesure des introductions successives d'allèles. Pour

les paramètres choisis les simulations aboutissent toutes à une diversité génétique maximale, c'est-à-dire une configuration avec tous les (64) génomes possibles présents dans la population. Lorsque $\alpha \geq 1$, la dynamique de la diversité génétique est très différente. En effet lorsque $\alpha \geq 1$, nous prouvons que les conditions d'invasion énoncées dans le Théorème 2.9 impliquent que la population à l'équilibre contient d'abord un allèle, puis toujours exactement deux allèles (lorsqu'un nouvel allèle apparaît dans la population qui en contient déjà deux, soit l'allèle mutant s'éteint soit il remplace un allèle initialement existant), et la distance génétique entre ces deux allèles est croissante. La population finit donc par contenir uniquement les deux génotypes les plus éloignés : $(0, ..., 0)$ et $(1, ..., 1)$.



(A) (B)

FIGURE 2.6 – **Évolution du nombre d'allèles maintenus dans la population, en supposant des mutations ponctuelles et une forme convexe de la fonction reliant la distance génétique et la fitness d'un couple d'allèles ($\alpha \leq 1$).** À partir d'une population initiale comportant deux allèles, nous induisons numériquement des mutations successives et suivons leur succès d'invasion au fil du temps. Le panneau (a) montre la distribution des allèles dans la population au fil du temps. Chaque couleur correspond à un allèle donné et la hauteur de la barre correspond au nombre d'individus porteurs de chaque allèle dans la population à un moment donné. Le panneau (b) indique le nombre d'allèles maintenus à l'équilibre après chaque mutation jusqu'à ce que le nombre total d'allèles soit atteint. Chaque ligne correspond à une simulation numérique différente ($n = 6$). Ici, $L = 6$ et $\alpha = 0,6$ de sorte qu'il y a $2^6 = 64$ allèles possibles.

2.3.4 Bilan des résultats

Dans cet ensemble de travaux nous avons étudié l'hétérogamie et l'homogamie, en utilisant des modèles individu-centrés multi-types et avec interaction. Nous avons créé et étudié des modèles les plus simples possibles, permettant de répondre aux questions qui nous intéressent. Pour cette

classe de modèles nous avons dans un premier temps donné des conditions explicites sous lesquels un mutant homogame a une probabilité d'invasion strictement positive, et nous avons caractérisé la probabilité et la durée de cette invasion. Ensuite nous avons montré comment l'homogamie couplée à une structuration spatiale et une migration liée à la recherche de partenaire sexuel peut générer une spéciation, ou l'arrêt d'un flux de gènes entre deux sous-populations. Enfin nous avons étudié le niveau de diversité génétique permis par l'homogamie et la construction progressive de cette diversité, par apparitions successives de mutants.

2.4 Éléments de preuves

Preuve du Théorème 2.3

La preuve du Théorème 2.3 consiste à étudier la dynamique de population en la découpant en trois phases : survie ou extinction de l'allèle mutant P, puis phase de croissance quasiment déterministe de la population, et enfin extinction de l'allèle résident p.

Survie du mutant Comme mentionné juste après la Proposition 2.2, tant que le nombre d'allèles P est petit, la dynamique du processus stochastique $(N_{AP}(t), N_{aP}(t))_{t\geq 0}$ peut être approximée par celle d'un processus de branchement multi-type surcritique, dont la matrice moyenne est donnée par l'Équation (2.5) et dont les probabilités d'extinction s_A et s_a partant respectivement d'un mutant de génotype AP ou d'un mutant de génotype aP satisfont donc le système d'équations

$$\begin{cases} b(1-s_A) + \bar{\beta}_{AA}(s_A^2 - s_A) + \bar{\beta}_{Aa}(s_A s_a - s_A) = 0 \\ b(1-s_a) + \bar{\beta}_{aa}(s_a^2 - s_a) + \bar{\beta}_{aA}(s_A s_a - s_a) = 0. \end{cases} \tag{2.26}$$

Cette approximation est justifiée en contrôlant la population résidente $(N_{Ap}(t), N_{ap}(t))$ jusqu'au temps $\inf(T_0^P, T_{\varepsilon^\xi})$ auquel le nombre d'allèles P atteint soit 0 soit $\varepsilon^\xi K$ où $\xi \in \{1/2, 1\}$. Nous montrons plus précisément (Proposition 3.1 de Coron *et al.* (2021)) qu'il existe une fonction η continue et nulle en 0 et une constante $\mathcal{A}_0 > 0$ telles que pour $\xi \in \{1/2, 1\}$,

$$\limsup_{K\to\infty} \left| \mathbb{P}\left(T_{\varepsilon^\xi} < T_0 \wedge R_{\mathcal{A}_0\varepsilon} \wedge U_{\varepsilon^{1/6}}, \left| \frac{T_{\varepsilon^\xi}}{\ln K} - \frac{1}{\lambda} \right| \leq \eta(\varepsilon) \middle| \mathbf{N}_P(0) = \mathbf{e}_\alpha \right) - (1 - q_\alpha) \right| = o_\varepsilon(1),$$

et

$$\limsup_{K\to\infty} \left| \mathbb{P}\left(T_0 < T_{\varepsilon^\xi} \wedge R_{\mathcal{A}_0\varepsilon} \wedge U_{\varepsilon^{1/6}} \middle| \mathbf{N}_P(0) = \mathbf{e}_\alpha \right) - q_\alpha \right| = o_\varepsilon(1), \tag{2.27}$$

où R_z (resp. U_z) est le temps au bout duquel le nombre de copies de l'allèle p s'éloigne de sa valeur initiale $K\frac{b-d}{c}$ de plus de z (resp. le temps au bout duquel la proportion d'individus de génotype Ap parmi les individus portant l'allèle p s'éloigne de sa valeur initiale ρ_A de plus de z) et par convention, $o_\varepsilon(1)$ tend vers 0 quand ε tend vers 0. Cette proposition est le résultat difficile de l'article. Pour la montrer nous utilisons la décomposition en semi-martingale du processus

$(N_{Ap}(t), N_{ap}(t), N_{AP}(t), N_{aP}(t))_{t\geq 0}$ et encadrons d'abord la proportion d'individus de génotype Ap parmi les individus portant l'allèle p (Lemme 3.3) puis la quantité de copies de l'allèle p (Lemme 3.4). Nous pouvons alors encadrer stochastiquement la population $(N_{AP}(t), N_{aP}(t))_{t\geq 0}$ par deux processus de branchement bi-types $N^{(\epsilon,-)}$ et $N^{(\epsilon,+)}$ tels que presque sûrement pour tout $t \leq T_{\varepsilon^\xi} \wedge T_0 \wedge R_{\mathcal{A}_0\varepsilon} \wedge U_{\varepsilon^{1/6}}$, $N_\alpha^{(\epsilon,-)}(t) \leq N_{\alpha P}(t) \leq N_\alpha^{(\epsilon,+)}(t)$ dont les probabilités d'extinction respectives $q_\alpha^{\epsilon,-}$ et $q_\alpha^{\epsilon,+}$ satisfont $0 \leq q_\alpha^{\epsilon,+} - q_\alpha^{\epsilon,-} \to 0$ quand ϵ tend vers 0 (Section 3.1.3 de Coron *et al.* (2021)).

Phase champ moyen Une fois que le mutant a atteint le niveau $\varepsilon^\xi K$, le Lemme 2.1 nous indique que le processus stochastique $(N_{Ap}(t), N_{ap}(t), N_{AP}(t), N_{aP}(t))_{t\geq 0}$ est bien approximé par le système dynamique décrit par l'Équation (2.11) dont les équilibres et les bassins d'attractions sont déterminés. Le problème est alors d'avoir une information sur la valeur "initiale" du couple (N_{AP}, N_{aP}) lorsque N_P dépasse $\varepsilon^\xi K$. Nous montrons que l'on peut trouver un temps assez court (inférieur à $T_{\sqrt{\varepsilon}}$) auquel les proportions $\frac{N_{AP}(t)}{N_{AP}(t)+N_{aP}(t)}$ et $\frac{N_{aP}(t)}{N_{AP}(t)+N_{aP}(t)}$ sont proches de celles données par le Théorème de Kesten-Stigum (Georgii et Baake (2003)), à savoir le vecteur propre à gauche de la matrice J (Équation (2.5)) associé à la valeur propre maximale λ.

Extinction du résident La dernière phase de la dynamique consiste à encadrer le temps d'extinction des individus de génotypes Ap, ap et aP et à contrôler la population d'individus AP pendant ce temps. Cette étape consiste à borner supérieurement le processus stochastique $(N_{Ap}(t), N_{ap}(t), N_{aP}(t))_{t\geq 0}$ par un processus de branchement multi-type sous critique et à utiliser des résultats classiques concernant ce type de processus stochastiques, que l'on peut trouver par exemple dans Athreya et Ney (1972).

Preuve des Théorèmes 2.5, 2.6 et 2.7

La preuve du Théorème 2.5 qui détermine les équilibres du système dynamique (2.11) et leur stabilité consiste à manipuler les équations stationnaires satisfaites par ces équilibres. La preuve du Théorème 2.6 est découpée en deux parties. Rappelons les définitions suivantes :

$$\mathcal{D} = \{\mathbf{z} \in \mathbb{R}_+^{\mathcal{E}}, z_{A,1} - z_{a,1} > 0, z_{a,2} - z_{A,2} > 0\},$$

et

$$\mathcal{K}_p = \left\{\mathbf{z} \in \mathcal{D},\ \{z_{A,1} + z_{a,1},\ z_{A,2} + z_{a,2}\} \in \left[\frac{b(\beta+1) - 2d - p}{2c}, \frac{2b\beta - 2d + p}{2c}\right]\right\}.$$

La première partie de la preuve du Théorème 2.6 consiste à prouver que n'importe quelle solution du Système (2.11) qui part dans $\mathcal{D}$ atteint puis reste dans le sous-ensemble $\mathcal{K}_p$ (ce sous-ensemble est stable). C'est l'objet du Lemme 2 dans Coron *et al.* (2018b). Sa preuve consiste à contrôler d'abord la taille totale de population puis les tailles respectives de population dans chaque patch. La seconde partie de la preuve du Théorème 2.6 consiste à exhiber une fonction de Lyapunov pour

le système dynamique (2.11) sur $\mathcal{K}_p$. Plus précisément nous considérons la fonction $V : \mathcal{D} \to \mathbb{R}$:

$$V(\mathbf{z}) = \ln\left(\frac{z_{A,1}+z_{a,1}}{z_{A,1}-z_{a,1}}\right) + \ln\left(\frac{z_{a,2}+z_{A,2}}{z_{a,2}-z_{A,2}}\right).$$

Nous prouvons (Lemme 3 de Coron *et al.* (2018b)) que la fonction V est une fonction de Lyapunov sur $\mathcal{K}_p$ si $p < p_0$. Ceci donne la convergence de toute solution du système dynamique (2.11) partant dans $\mathcal{D}$ vers l'équilibre approprié $(\zeta, 0, 0, \zeta)$. Une étude plus fine de la dynamique des différences $z_{A,1} - z_{a,1}$ et $z_{a,2} - z_{A,2}$ donne la convergence exponentiellement rapide une fois que l'ensemble $\mathcal{K}_p$ est atteint.

La preuve du Théorème 2.7 combine les résultats du Théorème 2.6 et une étude du processus stochastique $(\mathbf{Z}^K(t), t \geq 0)$ autour de l'équilibre $(\zeta, 0, 0, \zeta)$ quand K est grand. Plus précisément, notons $T_0^K = \inf\{t \geq 0, Z_{a,1}^K(t) + Z_{A,2}^K(t) = 0\}$ le temps auquel il n'y a plus d'individus de type a dans le patch 1 et plus d'individus de type A dans le patch 2 (qui est un état absorbant pour le processus stochastique considéré). Nous obtenons que ce temps est d'ordre $\log(K)/(b(\beta-1))$. Plus précisément nous prouvons (Proposition 2 de Coron *et al.* (2018b)) qu'il existe deux constantes positives ε_0 et C_0 telles que pour tout $\varepsilon \leq \varepsilon_0$, s'il existe $\eta \in]0, 1/2[$ tel que $\max(|z_{A,1}^0 - \zeta|, |z_{a,2}^0 - \zeta|) \leq \varepsilon$ et $\eta\varepsilon/2 \leq z_{a,1}^0, z_{A,2}^0 \leq \varepsilon/2$, alors

$$\begin{aligned}
&\text{pour tout } C > (b(\beta-1))^{-1} + C_0\varepsilon, \qquad \mathbb{P}(T_0^K \leq C\log(K)) \underset{K\to+\infty}{\to} 1,\\
&\text{pour tout } 0 \leq C < (b(\beta-1))^{-1} - C_0\varepsilon, \quad \mathbb{P}(T_0^K \leq C\log(K)) \underset{K\to+\infty}{\to} 0.
\end{aligned}$$

La preuve de cette proposition repose sur plusieurs arguments de couplage. La première étape consiste à prouver que les tailles de population $Z_{A,1}^K$ et $Z_{a,2}^K$ restent proches de ζ sur une longue échelle de temps. La deuxième étape consiste à coupler les processus $Z_{a,1}^K$ et $Z_{A,2}^K$ avec des processus de branchement sous-critiques dont les temps d'extinction sont connus, en s'appuyant sur des travaux antérieurs (Champagnat (2006), Théorème 3.c, et Freidlin *et al.* (1984), Chapitre 5).

Preuve des Théorèmes 2.8 et 2.9

La preuve du Théorème 2.8 consiste d'abord à donner, quand il existe, l'unique équilibre positif donné par l'Équation (2.24). Ensuite nous montrons par étude du Jacobien du système dynamique donné dans l'Équation (2.22) autour de l'unique équilibre, que cet équilibre est localement stable si et seulement si la matrice d'avantage sélectif M a une deuxième valeur propre strictement négative. Enfin nous montrons la stabilité globale de cet équilibre en montrant que la fonction

$$V(\mathbf{z}) := \sum_{\ell=1}^{k}\left(\frac{z_\ell}{z} - \frac{z_\ell^*}{z^*}\ln\left(\frac{z_\ell}{z}\right)\right) = 1 - \sum_{\ell=1}^{k}\frac{z_\ell^*}{z^*}\ln\left(\frac{z_\ell}{z}\right),$$

est une fonction de Lyapunov pour le système dynamique (2.22), si $z = \sum_l z_l$, $Z^* = (z_1^*, ..., z_k^*)$ et $z^* = \sum_l z_l^*$.

La preuve du Théorème 2.9 repose sur l'étude de la matrice Jacobienne du système dynamique de dimension $k+1$ autour du point $(Z^*, 0)$. Pour prouver que l'existence de l'équilibre à $k+1$ allèles entraîne sa stabilité globale nous utilisons le complément de Schur et le Théorème d'entrelacement des valeurs propres.

2.5 Perspectives

Plusieurs de mes perspectives de recherches sont connectées à ce sujet. Tout d'abord, nous cherchons actuellement, avec Manon Costa, Charline Smadi et Hélène Leman, à comprendre l'impact des structures familiales sur l'évolution génétique des populations. Nous considérons par exemple un modèle de Wright-Fisher dans lequel les individus forment des couples, qui seront choisis comme parents des individus de la génération suivante. Dans ce modèle les individus n'ont donc pas de demi-frères et sœurs, mais uniquement des frères et sœurs. Nous voulons étudier la composition génétique de la population dans ce modèle et notamment la comparer à celle obtenue pour un modèle de Wright-Fisher biparental classique. Ce projet permettrait de poursuivre nos investigations de l'impact des préférences d'appariement sur la composition génétique des populations. Dans une toute autre direction, avec Luis Almeida (CNRS, Sorbonne Université), je dirige la thèse de Léo Micollet, qui porte sur la modélisation mathématique du contrôle de populations d'insectes par lâcher de mâles stériles. Cette méthode de contrôle biologique appelée technique de l'insecte stérile a été utilisée en pratique pour lutter par exemple contre des ravageurs des élevages et des cultures comme la lucilie bouchère ou la mouche du cerisier, et consiste à relâcher régulièrement un grand nombre de mâles stériles avec lesquels les femelles présentes se reproduisent, produisant alors des œufs non viables. Pour modéliser cette méthode, Léo Micollet utilise des modèles individu-centrés multi-types, comme nous le faisons dans ce chapitre. Son but sera alors d'étudier la dynamique de cette population : ses différentes limites d'échelles (processus de branchement, système dynamique et processus de diffusion stochastique), sa probabilité d'extinction, le temps nécessaire à sa réémergence en cas d'immigration ou de taille de population très réduite, et, en lien avec le chapitre suivant, l'optimisation du coût de cette méthode de contrôle de population d'insectes, par suivi à l'aide de relevés de pièges permettant d'accéder à des données de comptage.

Chapitre 3

Suivi de la biodiversité à l'aide de données citoyennes

3.1 Introduction

Ce chapitre porte sur l'évaluation et le suivi de la biodiversité, à partir de jeux de données issus de programmes de sciences participatives. Ce travail, composé de deux articles (Giraud *et al.* (2015) et Coron *et al.* (2018a)) réalisés en collaboration avec Christophe Giraud (Laboratoire de Mathématiques d'Orsay), Clément Calenge (Office Français pour la Biodiversité) et Romain Julliard (CESCO, MNHN), est en continuité avec les sujets présentés dans les chapitres précédents, puisqu'il touche à la diversité génétique des populations, mais s'en détache nettement, par les approches mathématiques et le niveau de proximité avec les données.

Ce travail a deux motivations. La première est de réaliser des *cartes d'abondances relatives d'espèces*, c'est-à-dire, pour une espèce donnée, de comparer son abondance, soit le nombre d'individus de cette espèce, dans différentes zones de l'espace. Ces cartes et leurs dynamiques temporelles sont très importantes pour la société, notamment pour comprendre et anticiper l'extinction, le déplacement, ou encore l'émergence et l'invasion d'une espèce. Elles permettent aussi d'anticiper la réaction des espèces au réchauffement climatique ou à certaines modifications du territoire. Les scientifiques qui travaillent sur la biodiversité dépensent de ce fait beaucoup de temps, d'énergie et d'argent, à collecter des données (par exemple à partir de programmes de capture-marquage-recapture) permettant de créer de telles cartes. Néanmoins, depuis 40 ans environ, certains programmes, appelés programmes de sciences participatives, ou sciences citoyennes, permettent de faire appel aux compétences et à la motivation des citoyens pour récolter un grand nombre de données d'observation de la biodiversité. Ces programmes imposent en général un protocole très léger, de façon à maximiser la participation des citoyens. C'est là que réside notre deuxième motivation. Ces données citoyennes manquent de calibration, et en particulier résultent d'une intensité d'observation inconnue et très inégale spatialement et temporellement. Nous souhaitons comprendre comment elles peuvent, malgré ce biais, être utilisées pour améliorer notre

connaissance de la biodiversité.

Nous nous sommes pour cela placés dans une situation dans laquelle nous avons à notre disposition deux jeux de données d'observations : l'un issu d'un programme professionnel, et l'un issu d'un programme de sciences participatives. Cette situation est de plus en plus courante, et concerne même d'autres domaines scientifiques, comme la météorologie, le climat, l'étude de la qualité de l'air, des cours d'eau, etc... Plus concrètement, dans ce travail nous avons à notre disposition deux jeux de données d'observation d'oiseaux en Aquitaine. Le premier jeu de données est le résultat d'un protocole très précis imposé à des professionnels (plus de détails seront donnés dans la section suivante), tandis que le deuxième rassemble les observations de citoyens qui se sont simplement inscrits sur un site web et ont indiqué, s'ils le souhaitaient et quand ils le souhaitaient, leurs observations d'oiseaux, faites au cours de leur journée ou éventuellement lors de sorties dédiées à ces observations. L'objectif, à partir de ces données, est de réaliser des cartes d'abondances relatives d'espèces, c'est-à-dire, pour chaque espèce considérée, de fournir un estimateur du ratio du nombre d'individus de cette espèce, vivant dans deux zones différentes de l'espace. Estimer la carte d'abondance relative d'une espèce est possible en utilisant seulement le premier jeu de données, récolté selon un protocole calibré. Néanmoins les données issues d'observations citoyennes sont beaucoup plus nombreuses, et ont une couverture spatiale bien meilleure. Notre approche consiste à combiner ces deux jeux de données au travers d'un modèle probabiliste, de façon à bénéficier à la fois de la calibration apportée par les données professionnelles, et de l'abondance des données citoyennes. Nous obtenons alors que combiner les deux jeux de données permet une estimation plus précise des cartes d'abondances relatives, que le seul jeu de données professionnel (Théorème 3.1). Cette combinaison permet en outre de fournir des cartes d'abondances d'espèces pour certaines espèces qui ne sont pas observées dans le jeu de données professionnel, et d'estimer certains paramètres biologiques, comme les préférences des espèces considérées à différents types d'habitats (Section 3.4.2), qui sont des informations très importantes notamment pour prédire la réaction des espèces au changement climatique ou à certaines transformations du territoire.

Cette situation dans laquelle on dispose de plusieurs jeux de données d'observation d'une même réalité est de plus en plus courante, et je poursuis dans ce domaine par le co-encadrement de la thèse d'Emma Thulliez (INSA Rouen) qui porte sur l'évaluation de la qualité de l'air à partir de combinaison de mesures de concentration en dioxyde d'azote, réalisées par quelques stations fixes précises et un grand nombre de micro-capteurs de fiabilité bien moindre. Ce travail en cours est détaillé dans la section de perspectives 3.6.

3.2 Données

Comme mentionné précédemment, notre objectif est d'estimer des cartes d'abondances relatives d'espèces à partir de deux jeux de données d'observations d'oiseaux en Aquitaine : un jeu de données qui sera dit "standardisé" (car les observateurs ont dû suivre un protocole précis pour y participer), et un jeu de données qui sera dit "opportuniste" (car les observations rap-

portées peuvent avoir lieu lors de déplacements non nécessairement dévoués à ces observations). La qualité de nos estimations sera évaluée à l'aide d'un troisième jeu de données qui est aussi standardisé mais comporte moins de données que le premier. Je présente maintenant ces trois jeux de données.

Le premier jeu de données, standardisé, est fourni par l'Office Français pour la Biodiversité, et est appelé ACT (pour *Alaudidae, Columbidae, Turdidae*, qui sont les principaux clades d'oiseaux qui concernent ce programme ; le Tableau 3.1 présente la liste des espèces considérées). Dans ce programme, la région Aquitaine a été discrétisée en 64 quadrats, et dans chaque quadrat, 5 points espacés de 1km dans des habitats non urbains ont été définis. Ces points sont visibles dans la Figure 3.1(B). Chaque point a été visité deux fois, pendant exactement 10 minutes, le matin, et sous des conditions climatiques appropriées. Lors de chaque visite, l'espèce de chaque oiseau vu ou entendu a été enregistrée, et pour chaque espèce le maximum des deux comptes issus des deux visites a été retenu, selon des protocoles classiques dans le domaine de l'observation d'oiseaux. Les observateurs sont des professionnels, employés par l'Office Français pour la Biodiversité ou certaines associations de chasseurs. Entre 2008 et 2011, environ 9 500 observations d'oiseaux ont été rapportées.

Pour le jeu de données opportuniste, on utilise la base de données en ligne mise en place par la Ligue pour la Protection des Oiseaux. Chaque citoyen capable d'identifier des oiseaux peut s'enregistrer sur ce site web, et rapporter ses observations (ou certaines de ses observations) d'oiseaux, en mentionnant l'espèce, la date, l'heure et le lieu, à 500m près. Des centaines d'observateurs ont ainsi rapporté des centaines de milliers d'observations. Nous ignorons dans quel cadre ces observations ont été réalisées, et notamment la motivation des observateurs, s'ils rapportent certaines observations plutôt que d'autres ou non, ou encore leur temps passé à observer. Nous avons gardé uniquement les observations réalisées durant la même période de temps que celle du jeu de données standardisées, c'est-à-dire Avril à mi-Juin, entre 2008 et 2011. Ceci nous donne environ 115 000 observations de 34 espèces d'oiseaux (voir le Tableau 3.1), dont les lieux sont représentés dans la Figure 3.1 (B). Notons en particulier que seule une partie des espèces observées dans le programme de la LPO font partie du programme ACT présenté précédemment.

Le jeu de données que nous utiliserons pour évaluer la qualité de nos observations est fourni par le programme STOC (*Suivi temporel des oiseaux communs*, Jiguet *et al.* (2012)), un programme de surveillance des oiseaux nicheurs mis en place par le Muséum National d'Histoire Naturelle. Le protocole de ce programme est assez proche de celui du programme ACT, mais sans restriction sur les espèces d'intérêt. Ainsi les 34 espèces observées dans le jeu de données fourni par la Ligue pour la Protection des Oiseaux sont aussi présentes dans ce jeu de données. Entre 2008 et 2011, ce programme a donné lieu à 15 241 observations dans 251 points d'écoute, aussi dans un habitat non urbain.

TABLE 3.1 – Liste des 34 espèces d'oiseaux observées. Les 13 espèces suivies dans le programme ACT sont indiquées par une astérisque.

Latin name	Espèce	Latin name	Espèce
Aegithalos caudatus	Long-Tailed Tit	*Alauda arvensis**	Eurasian Skylark
*Alectoris rufa**	Red-Legged Partridge	*Carduelis carduelis*	European Goldfinch
Carduelis chloris	European Greenfinch	*Certhia brachydactyla*	Short-Toed Treecreeper
*Columba palumbus**	Common Wood Pigeon	*Coturnix coturnix**	Common Quail
Cuculus canorus	Common Cuckoo	*Cyanistes caeruleus*	Eurasian Blue Tit
Dendrocopos major	Great Spotted Woodpecker	*Erithacus rubecula*	European Robin
Fringilla coelebs	Common Chaffinch	*Garrulus glandarius**	Eurasian Jay
Hippolais polyglotta	Melodious Warbler	*Lullula arborea**	Woodlark
Luscinia megarhynchos	Common Nightingale	*Milvus migrans*	Black Kite
Parus major	Great Tit	*Passer domesticus*	House Sparrow
*Phasianus colchicus**	Common Pheasant	*Phoenicurus ochruros*	Black Redstar
Phylloscopus collybita	Common Chiffchaff	*Pica pica**	Eurasian Magpie
Pica viridis	Eurasian Green Woodpecker	*Sitta europaea*	Eurasian Nuthatch
*Streptopelia decaocto**	Eurasian Collared Dove	*Streptopelia turtur**	European Turtle Dove
Sylvia atricapilla	Eurasian Blackcap	*Troglodytes troglodytes*	Eurasian Wren
*Turdus merula**	Common Blackbird	*Turdus philomelos**	Song Thrush
*Turdus viscivorus**	Mistle Thrush	*Upupa epops*	Eurasian Hoopoe

3.3 Modèle

Nous divisons l'espace en J zones, ou sites, et nous voulons comparer, pour chacune des I espèces considérées, son abondance à deux sites différents. Pour cela, nous supposons que nous avons accès à deux jeux de données, indicés par k. On notera $k = 0$ pour le jeu de données standardisé, et $k = 1$ pour le jeu de données opportunistes. L'ensemble de ces jeux de données nous donne donc le nombre X_{ijk} d'observations d'individus de l'espèce $i \in [\![1, I]\!]$, dans la zone $j \in [\![1, J]\!]$, pour le jeu de données k. On modélise alors les comptages X_{ijk} par

$$X_{ijk} \sim \text{Poisson}(N_{ij}O_{ijk}), \quad \text{pour } i = 1, \dots, I, \ j = 1, \dots, J \text{ et } k = 0, 1,$$

où N_{ij} est le nombre d'individus de l'espèce i dans la zone j, et O_{ijk} modélise l'intensité résultant du protocole d'observation. La loi de Poisson est très classique dans l'analyse des données de comptage, et résulte de l'hypothèse qu'à chaque instant, chaque animal de chaque espèce est observé ou non, indépendant des autres animaux. Notre modèle néglige donc les interactions entre individus et entre espèces. Différentes pistes d'amélioration de ce travail sont évoquées dans la Section 3.6. Pour finir nous supposons pour des raisons d'identifiabilité que chaque zone a été visitée dans les deux jeux de données, et qu'au moins une espèce a été suivie dans les deux jeux de données.

3.3.1 Premier modèle

Dans un premier temps, nous supposons que l'intensité d'observation O_{ijk} est de la forme

$$O_{ijk} = P_{ik}E_{jk},$$

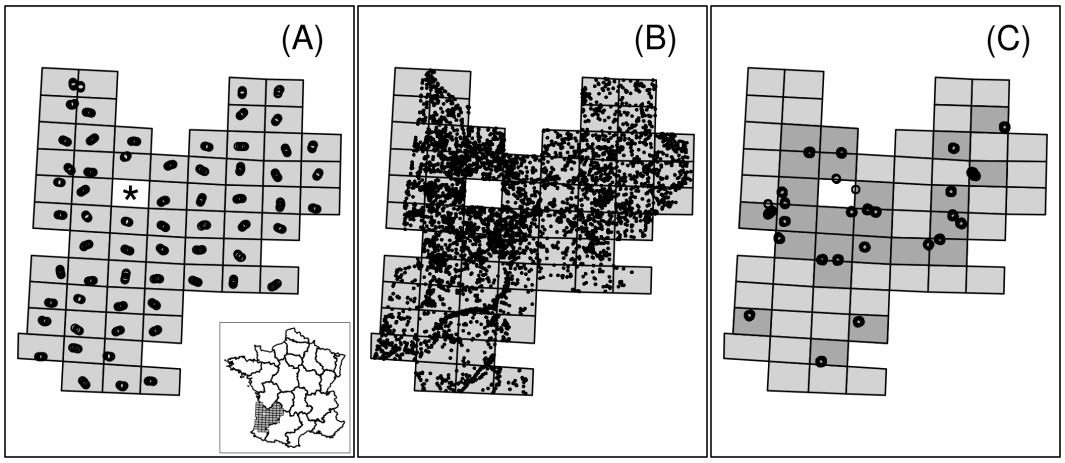


FIGURE 3.1 – Emplacements des observations pour les trois jeux de données disponibles. (A) Jeu de données ACT ; (B) Jeu de données LPO ; (C) Jeu de données STOC. L'espace est divisé en quadrats de taille 30×20 km, qui sont les sites, ou les zones dans notre analyse. Le quadrat contenant la métropole de Bordeaux (indiqué par une étoile dans la carte (A)) a été retiré de l'analyse.

où P_{ik} et E_{jk} sont des quantités vouées à modéliser les biais induits par la variabilité du processus d'observation. Notre modèle consiste donc à supposer que les impacts relatifs de l'espèce et de la zone sont indépendants. Un modèle plus complexe prenant en compte la division de l'espace en différents types d'habitats sera présenté dans la Section 3.3.2. Pour l'instant, notre modèle se réécrit donc :

$$X_{ijk} \sim \text{Poisson}(N_{ij} P_{ik} E_{jk}), \quad \text{pour } i = 1, \dots, I, \ j = 1, \dots, J \text{ et } k = 0, 1, \tag{3.1}$$

Les paramètres P_{ik} peuvent être interprétés comme la probabilité de détection et de report d'une observation d'un individu de l'espèce i, pour le jeu de données k. Les paramètres E_{jk}, que l'on appellera intensité d'observation dans la zone j pour le jeu de données k, représente l'impact de la variabilité de l'effort (typiquement le temps passé à observer, le nombre de sorties, le nombre d'observateurs, la variabilité des conditions d'observations...) dans la zone j pour le jeu de données k. Rappelons que nous nous intéressons au cas où le premier jeu de données est standardisé, tandis que le deuxième est opportuniste. Cela peut s'interpréter mathématiquement par le fait que l'intensité d'observation est connue (à une constante multiplicative près) pour le jeu de données standardisé ($k = 0$), et peut être supposée très grande pour le jeu de données opportuniste ($k = 1$). Typiquement pour le jeux de données standardisé fourni par l'Office français pour la biodiversité et présenté dans la Section 3.2 nous supposons que l'effort d'observation est le même dans chaque zone j, donc E_{j0} ne dépend pas de j. En revanche les intensités d'observations E_{j1} sont supposées grandes mais inconnues.

3.3.2 Raffinement : ajout d'une structure

Dans cette section, nous proposons un modèle plus général, qui prend en compte l'impact de covariables, comme le type d'environnement (urbain, forestier, agricole, ...), la densité de population humaine, l'altitude,..., à la fois sur l'abondance de chaque espèce et sur le processus d'observation. Pour plus de simplicité, nous supposerons par la suite que ces covariables sont des types d'environnement, mais notre approche peut être plus générale. Dans notre travail, l'habitat associé à chaque observation pourra être inconnu, il sera traité comme une variable latente, dont la loi est caractérisée par des paramètres appelés préférences d'habitat, que nous chercherons à estimer. Les différents types d'habitat sont indicés par $h \in \{1, 2, ..., H\}$.

Espèces : abondances et préférence d'habitat Comme dans la Section 3.3.1, on note N_{ij} le nombre d'individus de l'espèce i dans la zone j. On suppose alors que la densité de l'espèce i au point x de la zone j est donnée par

$$\frac{N_{ij} S_{ih(x)}}{\sum_{h'} S_{ih'} V_{h'j}^{zone}},$$

où V_{hj}^{zone} est l'aire occupée par l'habitat h dans la zone j et $h(x)$ est le type d'habitat au point x. Les paramètres S_{ih} peuvent donc être vus comme la préférence de l'espèce i pour l'habitat h, qui a un intérêt applicatif fort, et que nous souhaitons estimer.

Observations rapportées Comme dans la Section 3.3.1, on indice le jeu de données standardisé par $k = 0$ et le jeu de données opportuniste par $k = 1$. Chaque zone j est alors divisée en plusieurs cellules indicées par c, telles que l'on connaît la cellule dans laquelle chaque observation a eu lieu. Cette cellule pouvant typiquement être une commune, un quartier, ou un groupe de communes. Notons que le découpage en cellules peut dépendre du jeu de données et que dans chaque jeu de données seule une fraction des cellules a été visitée au moins une fois par les observateurs. Pour une cellule c visitée dans le jeu de données k, on note X_{ick} le nombre d'observations rapportées, d'un individu de l'espèce i. Comme pour le modèle précédent, on note E_{ck} l'intensité d'observation dans la cellule c pour le jeu de données k, et on note P_{ik} la probabilité de détection et rapport d'une observation de l'espèce i pour le jeu de données k. Pour le jeu de données k, on modélise l'intensité d'observation au point x dans la cellule c par

$$\frac{q_{h(x)k} E_{ck}}{\sum_{h'} q_{h'k} V_{h'c}^{cell}},$$

où V_{hc}^{cell} est l'aire (connue) de la cellule c couverte par l'habitat h et $q_{hk} \in [0, 1]$ modélise la préférence des observateurs pour l'habitat h, pour le jeu de données k. On a alors

$$X_{ick} \sim \mathcal{P}\text{oisson}\left(N_{ij} E_{ck} P_{ik} \sum_{h} \frac{q_{hk}}{\sum_{h'} q_{h'k} V_{h'c}} \times \frac{S_{ih}}{\sum_{h'} S_{ih'} V_{h'j}} V_{hc} \right). \tag{3.2}$$

Rappelons que les aires V_{hj}^{zone} et V_{hc}^{cell} sont connues dans ce modèle. Comme précédemment on supposera naturellement que pour le jeu de données standardisées les intensités d'observations E_{c0} sont connues (à une constante multiplicative près), et nous supposons par ailleurs que l'habitat associé à chaque observation standardisée est soit connu, soit satisfait que les ratios q_{h0}/q_{10} sont connus pour tout h. Les autres paramètres sont inconnus.

Notre modèle consiste à supposer une différence d'échelle spatiale entre les individus observés et les observateurs : les oiseaux choisissent leur position (ou leur habitat) à l'échelle de la région, tandis que les observateurs choisissent la position de leur observation à l'échelle de la cellule (qui sera typiquement de l'échelle de la commune). Enfin, supposer que les observateurs comme les individus observés n'ont pas de préférences particulières à certains types d'habitats revient à poser $q_{hk} = S_{ih} = 1$ pour tout h, auquel cas on obtient que le nombre X_{ick} d'observations de l'espèce i dans la cellule c de la zone j pour le jeu de données k suivra le premier modèle (3.1). Le modèle que nous venons de proposer consiste donc bien en un raffinement du premier modèle.

3.4 Résultats : théorie, et application aux données

3.4.1 Résultats théoriques

Identifiabilité et estimation des paramètres Pour les deux modèles (3.1) et (3.2), nous pouvons prouver par des changements de variables adéquats l'identifiabilité des abondances relatives N_{ij}/N_{i1} ainsi que des préférences à l'habitat S_{ih}, qui sont les deux quantités qui nous intéressent, à partir des observations.

Plus précisément pour le modèle (3.1), le changement de variables

$$\widetilde{N}_{ij} = N_{ij} P_{i1} E_{10} \frac{P_{10}}{P_{11}}, \qquad \widetilde{E}_{jk} = \frac{E_{jk}}{E_{10}} \times \frac{P_{1k}}{P_{10}}, \qquad \text{et} \qquad \widetilde{P}_{ik} = \frac{P_{ik}}{P_{i1}} \times \frac{P_{11}}{P_{1k}}.$$

est tel que pour toute espèce i, toute zone j, tout jeu de données k, $\widetilde{N}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik} = N_{ij}E_{jk}P_{ik}$, $\widetilde{N}_{ij}/\widetilde{N}_{i1} = N_{ij}/N_{i1}$, et $\widetilde{E}_{j0} = E_{j0}/E_{10}$ est connu. Alors en posant $n_{ij} = \log(\widetilde{N}_{ij})$, $e_{jk} = \log(\widetilde{E}_{jk})$ et $p_{ik} = \log(\widetilde{P}_{ik})$, le modèle (3.1) peut être vu comme un modèle linéaire généralisé :

$$X_{ijk} \sim \text{Poisson}(\lambda_{ijk}), \quad \text{avec } \log(\lambda_{ijk}) = n_{ij} + e_{jk} + p_{ik}, \tag{3.3}$$

où $e_{j0} = \log \widetilde{E}_{j0}$ est connu, $p_{i1} = 0$ pour tout i, et $p_{10} = 0$. On note alors l'estimateur par maximum de vraisemblance $(\widehat{N}_{ij}, \widehat{E}_{jk}, \widehat{P}_{ik})$ des paramètres $\widetilde{N}_{ij}$, $\widetilde{E}_{jk}$ et $\widetilde{P}_{ik}$, qui peut être obtenu en pratique en utilisant la commande glm dans R.

Le modèle présenté dans la Section 3.3.2, qui intègre une structure en habitat, contient des non-linéarités qui exigent une autre approche d'estimation car la vraisemblance de ses paramètres ne peut pas être maximisée de cette façon. Nous choisissons une approche d'estimation bayésienne implémentée par l'échantillonneur de Gibbs *JAGS* (Plummer (2003)). Nous appelons ce programme dans *R* (R Core Team (2014)) en utilisant le package *rjags* (Plummer (2014)).

Nous choisissons des priors non informatifs pour les différents paramètres à estimer et l'échantillonneur *JAGS* fournit, en sélectionnant les données simulées les plus proches des données réelles, des échantillons distribués selon la distribution a posteriori des paramètres.

Amélioration de la précision des cartes (sans structure spatiale) Notre premier résultat montre, pour le premier modèle (3.1), que les estimations des abondances relatives d'espèces obtenues en combinant le jeu de données standardisé (pour lequel les ratios d'efforts d'observation $E_{j0}/E_{j'0}$ sont connus) et le jeu de données opportuniste (pour lequel les efforts E_{j1} peuvent être supposés grands) sont asymptotiquement meilleures que celles obtenues en utilisant uniquement le jeu de données standardisé. Notre résultat permet même de quantifier la réduction de variance obtenue grâce à la combinaison des données.

Théorème 3.1. *Sous le Modèle (3.1), posons $\widehat{N}_{ij}$ et $\widehat{N}^0_{ij}$ les estimateurs de maximum de vraisemblance du paramètre $\tilde{N}_{ij}$ obtenus en utilisant respectivement les deux jeux de données (standardisé et opportuniste), ou le jeu de données standardisé uniquement. On a*

(*i*)

$$\lim_{E_{j1}\to\infty} \mathbb{E}\left(\frac{\widehat{N}_{ij}}{\widehat{N}_{i1}}\right) = \mathbb{E}\left(\frac{\widehat{N}^0_{ij}}{\widehat{N}^0_{i1}}\right) = \frac{\widetilde{N}_{ij}}{\widetilde{N}_{i1}} = \frac{N_{ij}}{N_{i1}} \tag{3.4}$$

(*ii*)

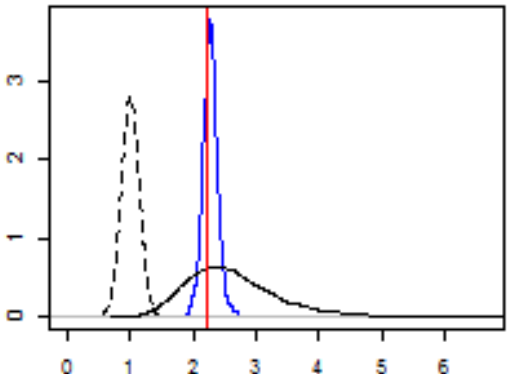
$$\lim_{E_{j1}\to\infty} Var(\widehat{N}_{ij}) = Var(\widehat{N}^0_{ij}) \times \frac{P_{i0}N_{ij}}{\sum_l P_{l0}N_{lj}}\,. \tag{3.5}$$

Ce résultat nous dit que lorsque l'effort d'observation ayant généré les données opportunistes est très grand (ce qui est ce que l'on attend de ce type de jeux de données), alors l'ajouter au jeu de données standardisées pour estimer les abondances relatives d'espèces permet de faire décroître la variance des estimations, en la multipliant par un facteur $\frac{P_{i0}N_{ij}}{\sum_l P_{l0}N_{lj}}$. Cette réduction de variance est en particulier importante pour les espèces rares (i.e. lorsque N_{ij} est petit), difficiles à détecter (i.e. lorsque P_{i0} est petit), ou lorsque le nombre I d'espèces suivies est grand.

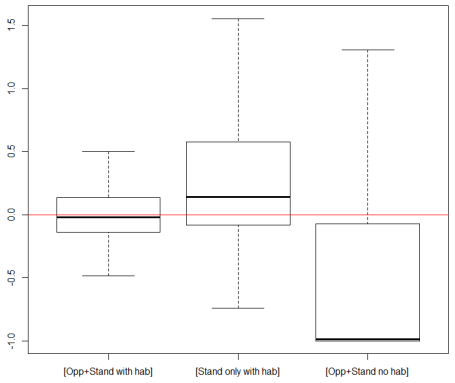
Amélioration de la précision des cartes (avec structure spatiale) Pour le Modèle (3.2) qui prend en compte une structuration de l'espace en plusieurs types d'habitats, nous évaluons la performance de notre approche de combinaison de jeux de données en utilisant des données simulées. Plus précisément nous fixons tous les paramètres nécessaires (les abondances de population N_{ij}, les efforts E_{ck}, les probabilités P_{ik}, et les préférences q_{hk} et S_{ih}, puis nous générons deux jeux de données, suivant le Modèle (3.2). Nous estimons alors les paramètres d'intérêt (notamment les abondances relatives N_{ij}/N_{i1} et les préférences S_{ih}/S_{i1}), et nous comparons les distribution postérieures obtenues, aux valeurs fixées pour ces paramètres. Plus précisément nous réalisons ces estimations pour trois situations : (i) en utilisant uniquement les données standardisées et le modèle (3.2) qui a été utilisé pour générer les jeux de données ([Stand only with hab]), (ii) en utilisant les deux jeux de données et toujours le modèle (3.2) ([Opp+Stand with hab]), (iii) en

utilisant les deux jeux de données mais en faisant nos estimations en supposant que les données sont générées selon le premier Modèle (3.1) ([Opp+Stand no hab]).

Les Figures 3.2A et 3.2B montrent, comme prouvé dans le Théorème 3.1 en l'absence de structuration en habitats, que l'estimation obtenue en combinant les deux jeux de données est plus précise que celle obtenue en utilisant uniquement le jeu de données standardisé. Elles illustrent aussi le fait que, sans surprise, le fait de négliger les préférences pour différents types d'habitats mène à des estimations biaisées des abondances relatives d'espèces. Notons que pour la Figure 3.2B nous choisissons comme abondance relative estimée $\hat{N}_{ij}/\hat{N}_{i1}$ la moyenne de la distribution postérieure de N_{ij}/N_{i1}.



(A) Exemple de distribution a posteriori d'une abondance relative d'espèce N_{i2}/N_{i1} pour une espèce i fixée, estimée avec la procédure [Stand only with hab] (en noir), [Opp+Stand with hab] (en bleu) et [Opp+Stand no hab] (en pointillé). La valeur fixée (que l'on cherche à retrouver) est en rouge.



(B) Boxplots des différences relatives $\frac{(\hat{N}_{ij}^{[model]}/\hat{N}_{i1}^{[model]})-(N_{ij}/N_{i1})}{N_{ij}/N_{i1}}$ entre abondances estimées et "réelles", en utilisant les deux jeux de données (Opp+Stand) ou en utilisant uniquement le jeu de données standardisé (Stand only), et en prenant en compte, ou non, la structuration en habitat.

FIGURE 3.2

3.4.2 Application aux données étudiées

En pratique, les "zones" de notre modèle, indicées par j, seront les 63 quadrats définis dans le cadre du programme ACT (Fig. 3.1(A)). Le quadrat contenant la zone de Bordeaux a été enlevé car nous avons supposé et observé que le comportement des observateurs opportunistes et la structuration de l'espace en différents habitats étaient trop différents dans ce quadrat, par rapport aux autres. Pour les deux jeux de données ACT et STOC, nous prenons comme intensité d'observation dans la zone j simplement le nombre de visites dans cette zone entre les années 2008 et 2011, puisque toutes les visites doivent avoir la même durée. L'habitat a été défini en utilisant l'occupation du sol, rendue disponible par Corine Land Cover. Plus précisément nous

avons retenu 7 catégories d'habitat permettant à la fois l'identifiabilité du modèle et une étude pertinente du comportement des espèces considérées : zone urbaine, agriculture intensive, paysage naturel ouvert, terres agricoles, forêt de conifères, forêts d'arbres à feuilles caduques, et forêts mixtes.

Réduction de variance Le Tableau 3.2 illustre la réduction de variance énoncée dans le Théorème 3.1, pour les jeux de données présentés dans la Section 3.2. Pour remplir ce tableau nous calculons pour chaque espèce la corrélation entre les abondances relatives estimées pour toutes les zones par notre approche (soit avec le seul jeu de données standardisé ACT, soit en combinant le jeu de données standardisées ACT et le jeu de données opportunistes LPO) et les abondances relatives estimées en utilisant le jeu de données standardisé de référence, STOC. Les deux premières lignes du Tableau 3.2 donnent la médiane de ces corrélations obtenues pour chaque espèce, ainsi que leurs premier et dernier quartiles. Nous obtenons que ces médianes et quartiles sont plus élevés en combinant les deux jeux de données ACT et LPO, ce qui illustre le fait que les estimations obtenues en combinant les deux jeux de données sont plus précises que celles obtenues en utilisant seulement le jeu de données standardisé.

Il s'avère que le jeu de données standardisé que nous utilisons est très riche : du fait des dix visites par zone et par année il est le résultat d'une intensité d'observation très élevée, ce qui n'est pas a priori nécessaire pour notre approche. Nous avons donc dans un deuxième temps étudié dans quelle mesure l'amélioration des estimations apportée par la combinaison des jeux de données se maintient lorsque le jeu de données standardisé est moins riche. Nous avons pour cela réduit artificiellement le jeu de données ACT en sélectionnant aléatoirement une seule visite par zone et en réduisant les données du jeu de données ACT à celles obtenues lors de ces visites sélectionnées. La taille du jeu de données ACT est divisée environ par 18, suite à cette opération. Les deux dernières lignes du Tableau 3.2 comportent les mêmes quantités que les deux premières lignes, mais en remplaçant le jeu de données standardisé ACT par ce nouveau jeu de données artificiellement réduit et dont les capacités prédictives deviennent alors faibles (corrélation quasiment nulle avec les estimations produites par le jeu de données de référence STOC). Nous obtenons en revanche que la qualité des estimations obtenues en combinant les deux jeux de données a très peu diminué. Ces lignes illustrent en pratique l'intérêt de notre approche, déjà énoncé dans le Théorème 3.1 : combiner un jeu de données standardisées de taille très faible avec un jeu de données opportuniste permet d'estimer les abondances relatives de plusieurs espèces de façon plus satisfaisante et à moindre coût. La combinaison des jeux de données permet en outre de donner des estimations pour les abondances relatives d'espèces qui ne sont pas observées dans le jeu de données standardisé (deuxième colonne du Tableau 3.2). Ces estimations sont aussi de meilleure qualité que celles obtenues par le seul jeu de données standardisé, pour les espèces qu'il contient.

Des résultats similaires sont donnés dans le Tableau 3.3 pour le modèle avec structure spatiale (3.2). Notons que les résultats des Tableaux 3.2 et 3.3 diffèrent légèrement pour le modèle (3.1), ce qui s'explique par des méthodes d'estimation différentes : maximum de vraisemblance pour

le Tableau 3.2 et estimation Bayesienne avec prior non informatif pour le Tableau 3.3. Comme précédemment nous obtenons que combiner les deux jeux de données et prendre en compte la structure spatiale en habitat permet une amélioration des estimations.

Jeux de données utilisés	Dans ACT	Pas dans ACT
Standardisé	0.27 (0.13 – 0.49)	—
Standardisé + opportuniste	0.55 (0.38 – 0.68)	0.35 (0.19 – 0.47)
Standardisé réduit	0.06 (-0.07 – 0.23)	—
Standardisé réduit + opportuniste	0.54 (0.25 – 0.61)	0.28 (0.08 – 0.40)

TABLE 3.2 – Médiane (et premier et dernier quartiles entre parenthèses) des corrélations, pour chaque espèce, entre abondances relatives estimées en utilisant notre approche d'une part et le jeu de données de référence (STOC) d'autre part.

Données et modèle	Dans ACT	Pas dans ACT
[Opp+Stand avec hab]	0.49 (0.30–0.54)	0.39 (0.12–0.54)
[Stand only avec hab]	0.29 (0.03–0.46)	–
[Opp+Stand sans hab]	0.44 (0.32–0.68)	0.31 (0.19–0.42)

TABLE 3.3 – Médiane des corrélations (ainsi que premier et dernier quartiles) entre les estimations d'abondances relatives obtenues pour chaque espèce grâce aux observations STOC seules d'une part et grâce à différentes approches d'autre part : [Opp+Stand avec hab] correspond à la prise en compte de l'habitat et la combinaison des jeux de données, [Stand only avec hab] correspond à la prise en compte de l'habitat et l'utilisation du seul jeu de données standardisé, et [Opp+Stand sans hab] correspond à la combinaison des jeux de données mais sans prise en compte de l'habitat.

Cartes d'abondances relatives Cette section donne quelques résultats écologiques qui sont un autre fruit de notre travail. La Figure 3.3 donne la carte d'abondance de la sitelle torchepot, en utilisant les modèles avec et sans prise en compte de l'habitat. Une telle carte peut bien sûr être donnée pour chaque espèce du jeu de données. Remarquons aussi que notre approche pourrait aussi donner l'évolution de cette carte au cours du temps, en remplaçant les zones spatiales par des couples "zone-année" par exemple. Cette analyse, couplée à l'estimation des préférences de chaque espèce à chaque type d'habitat (présentée dans le prochain paragraphe), serait pertinente dans le cadre de la prédiction de la réaction des espèces au réchauffement climatique. Elle nécessiterait toutefois, pour plus d'intérêt, d'avoir des données couvrant un plus grand nombre d'années.

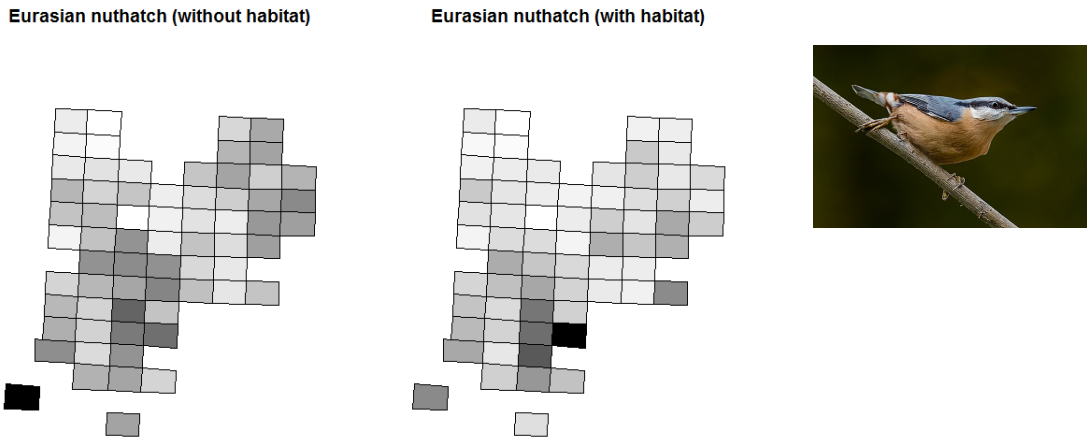


FIGURE 3.3 – Cartes d'abondances relatives de la sitelle torchepot avec ou sans prise en compte de l'habitat. Pour chaque quadrat, le niveau de gris matérialise une version renormalisée entre 0 et 1 de la quantité $\hat{N}_{ij}^{model}$ (i.e. ces niveaux de gris vont de 0 (abondance la plus faible, quadrat blanc) à 1 (abondance la plus élevée, quadrat noir).

Estimation des préférences d'habitat Comme mentionné précédemment, un produit utile de notre approche est l'estimation des préférences $(S_{ih})_{1\leq h\leq H}$ de chaque espèce i pour les différents types d'habitats, qui sont des quantités très importantes pour les écologues et qui requièrent habituellement de gros efforts pour être estimées (Lele *et al.* (2013); Boyce et McDonald (1999)). Dans la Figure 3.4 nous donnons à titre d'exemple les préférences que nous obtenons pour le pic épeiche. Cette espèce est connue pour préférer les habitats forestiers, ce qui se retrouve dans nos estimations.

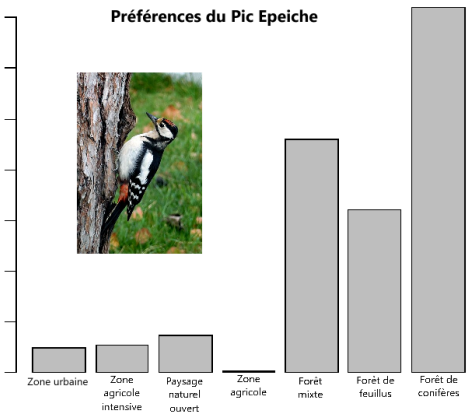


FIGURE 3.4 – Préférences (relatives) du pic épeiche pour les différents types d'habitats considérés.

3.5 Éléments de preuves

La preuve du Théorème 3.1 repose sur l'analyse du modèle linéaire (3.3). Tout d'abord nous montrons l'identifiabilité des paramètres de ce modèle en montrant que sa matrice de régression a un noyau de dimension $I + J + 1$ tel que les contraintes données juste après l'Équation (3.3) (qui correspondent à rappeler que l'on dispose d'un jeu de données standardisé pour lequel les efforts d'observation sont connus, à constante multiplicative près) permettent l'identifiabilité des abondances relatives étudiées. La suite de la preuve consiste à exprimer l'estimateur de maximum de vraisemblance de ces abondances relatives et à en étudier l'espérance et la variance, notamment lorsque l'effort correspondant au jeu de données opportuniste tend vers l'infini.

3.6 Perspectives

Améliorations du modèle Le modèle que nous considérons suppose que les individus observés (ainsi que les observateurs) se comportent indépendamment les uns des autres. Il néglige également le fait que les erreurs d'identifications peuvent créer des interactions entre les nombres d'observations d'oiseaux de différentes espèces (plus fréquemment confondues l'une avec l'autre, typiquement), et que ces interactions peuvent varier d'une zone à l'autre. Ces deux limites du modèle peuvent respectivement être résolues en remplaçant la distribution de Poisson par une autre distribution et en faisant dépendre la probabilité de détection d'une espèce de l'abondance des autres espèces. Ce nouveau modèle serait plus complexe à étudier mais conduirait à des résultats plus réalistes et à une analyse intéressante des interactions entre individus et entre espèces.

Combinaison de données pour la qualité de l'air Comme mentionné dans l'introduction de ce chapitre, la situation dans laquelle les scientifiques ont accès à différents ensembles de données mesurant la même quantité ou étudiant le même phénomène est désormais très courante. La combinaison de jeux de données au travers de modèles probabilistes dont les paramètres peuvent être inférés par des méthodes statistiques est donc une question mathématique intéressante ayant de forts enjeux applicatifs. Je co-encadre actuellement la thèse de doctorat d'Emma Thulliez (INSA Rouen), avec Bruno Portier, et en collaboration d'une part avec Jean-Michel Poggi (Université Paris-Saclay) et d'autre part avec ATMO Normandie qui est une association chargée par l'État d'évaluer la qualité de l'air en Normandie. Un des objectifs de cette thèse est de produire des cartes de concentration de certains polluants dans l'air (comme le NO_2 ou les particules fines). Pour ce faire, en prenant l'exemple du NO_2 (dioxyde d'azote), nous avons accès à trois jeux de données :

— Des ensembles de cartes de concentration en NO_2 sur une zone donnée à différents instants, qui sont des sorties de modèles physico-chimiques, prenant en compte l'intensité annuelle du trafic, les émissions des entreprises, la forme des routes, ainsi que certaines données météorologiques (la température, la vitesse du vent, l'humidité, etc...). Un exemple de telle

carte, issue du modèle SIRANE (Soulhac *et al.* (2017)), est donné dans la Figure 3.5, et nous pouvons typiquement avoir accès à une carte par heure, sur une période donnée. Notons que ces modèles physico-chimiques sont régulièrement améliorés, pour prendre en compte de nouveaux phénomènes. Ils négligent par ailleurs certaines spécificités locales, comme l'altitude, la pente des rues, les travaux de voirie, etc...

— Des mesures de concentration en NO_2 dans l'air réalisées par quelques (4, pour la métropole de Rouen) stations de références, qui sont supposées fournir des mesures précises et non biaisées de la vraie concentration, que nous cherchons à estimer. Ces stations fournissent aussi des mesures de concentrations d'autres polluants ainsi que des mesures météorologiques.

— Des mesures de concentration en NO_2 réalisées par un grand nombre (jusqu'à 70, pour la métropole rouennaise) de micro-capteurs qui mesurent la même concentration mais en utilisant une technologie très différente de celle des stations fixes. Ces micro-capteurs sont beaucoup moins chers, mais leurs mesures sont moins précises que celles des stations fixes, et elles sont également biaisées. La Figure 3.6 montre des mesures réalisées par un micro-capteur durant un mois, ainsi que les estimations fournies par la carte SIRANE à l'emplacement de ce micro-capteur. Ces appareils fournissent également des mesures pour plusieurs autres quantités, telles que la pression, la température, la vitesse du vent, ... Nous pourrons aussi par la suite utiliser des mesures réalisées par des micro-capteurs mobiles qui ont été installés récemment sur des bus de la métropole. Notons que la Figure 3.6 montre un changement de comportement du micro-capteur un peu après le milieu de la période considérée. Ce changement indique que des approches consistant à calibrer les micro-capteurs en les accolant temporairement à des stations fixes, qui sont souvent utilisées, ont peu de chances d'être performantes dans ce contexte.

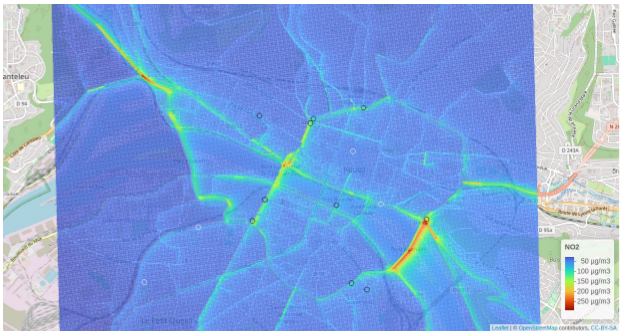


FIGURE 3.5 – Un exemple de carte de concentration en NO_2 issue du modèle SIRANE.

Notre approche consiste à supposer que les résultats du modèle physico-chimique ne peuvent pas être exacts, car ce modèle fait certaines hypothèses, comme une altitude constante ou une occupation du sol constante. Ils négligent généralement la présence de parcs et n'ont pas accès à l'intensité précise du trafic, et à la présence de travaux de voirie, alors que ces caractéristiques

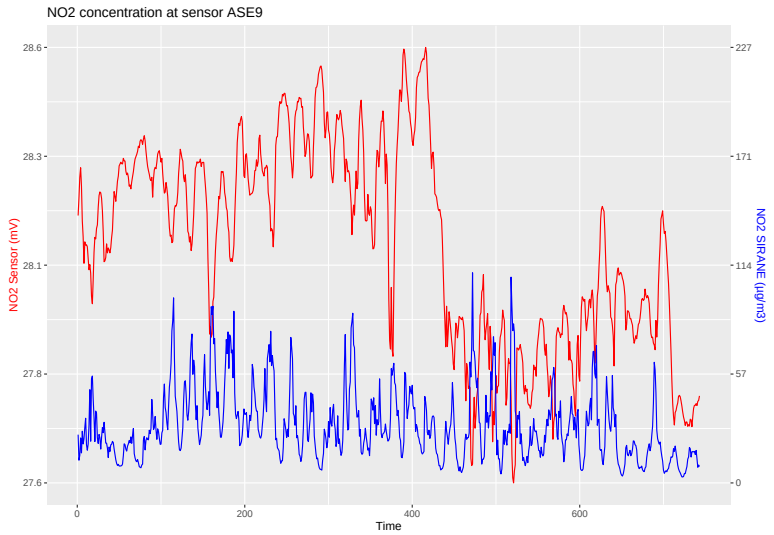


FIGURE 3.6

spatio-temporelles jouent un rôle dans la qualité de l'air. Nous modélisons donc les biais de ce modèle physico-chimique et estimons les paramètres de ce biais en utilisant les stations fixes disponibles et les mesures des micro-capteurs. Cette combinaison de données de qualités différentes est différente de celle rencontrée pour l'utilisation de programmes de science citoyenne dans l'évaluation de l'abondance des espèces, mais nous traitons ces deux questions en utilisant une approche de modélisation probabiliste intégrative similaire.

Retour à la génétique Avec Sophie Donnet (INRAE), Raphaël Leblois (INRAE), Miguel De Navascués (INRAE), Julien Stoehr (CEREMADE, Université Paris Dauphine), nous souhaitons développer des méthodes statistiques permettant de combiner au travers de modèles probabilistes des données génétiques et des données de comptage, de façon à améliorer l'estimation de paramètres démographiques qui pourrait être faite en utilisant seulement l'un des jeux de données. Ces données pourront être des données issues de programme d'observations professionnels ou citoyens, comme ceux que j'ai présentés dans la Section 3.2 de ce chapitre, ou bien provenir de protocoles de capture-marquage-recapture par exemple. Les modèles de dynamique de population qui pourront être utilisés dans ce projet seront par exemple des modèles de naissance et mort avec interaction du type de ceux étudiés dans le Chapitre 2 ou éventuellement des modèles de génétique de population comme celui présenté dans le Chapitre 1. Pour ce projet nous avons proposé un post-doctorat qui sera réalisé par Lucas Rey (Université Paris-Dauphine), et des applications à l'étude de la dynamique d'espèces de ravageurs de cultures seront abordées dans ce cadre. Cette dernière perspective pourra aussi rejoindre la perspective mentionnée en fin de Chapitre 2 sur le contrôle de population d'insectes par technique de l'insecte stérile. En effet les ravageurs de culture comme Drosophile Suzukii (la mouche du cerisier) sont des exemples d'es-

pèces pour lesquelles à la fois des données de comptage et des données de séquençage génétique sont susceptibles d'être disponibles. Comme ces données présentent à la fois des coûts et des intérêts différents, en termes d'inférence démographique, la question d'un éventuel arbitrage entre les récoltes de ces deux types de données peut être posée. Pour finir ce dernier sujet rejoint le sujet de la thèse d'Arnaud Becheler (Becheler (2018)) que j'ai co-encadrée avec Stéphane Dupas (IRD, Gif-s/-Yvette) et qui portait sur le développement et l'implémentation d'un modèle pour la dynamique démo-génétique du frelon asiatique, dans un paysage structuré.

Conclusion et bilan des perspectives

Durant les dernières années je me suis intéressée, de façon principale, à la modélisation et l'étude mathématique de l'évolution génétique des populations à reproduction sexuée et de la biodiversité ainsi qu'à la combinaison de jeux de données. J'ai essayé de trouver un équilibre entre la simplicité des modèles, qui permet leur étude mathématique et la compréhension essentielle des phénomènes étudiés, et la complexité des modèles qui permet leur confrontation à des données de façon pertinente. J'ai hâte de poursuivre mon travail dans ces deux directions et de découvrir de nouveaux sujets, de nouveaux enjeux, de nouvelles collaborations.

Les perspectives que j'ai développées à la fin de chaque chapitre portent d'abord sur l'étude de la proportion de génome transmis dans les populations biparentales, dans la lignée des travaux que j'ai présentés dans le Chapitre 1. J'aimerais notamment enrichir ce travail par l'étude de l'impact de différentes formes de sélection sur la proportion de génome transmis par un individu, mais aussi relier ce travail à des approches beaucoup plus appliquées qui consistent à estimer des paramètres d'histoire démographique à partir de données génétiques. Ensuite, les travaux que j'ai présentés dans le Chapitre 2 et qui portent notamment sur l'étude des limites d'échelles et du contrôle des processus de naissance et mort avec interactions trouveront un débouché naturel dans l'étude par Léo Micollet dans le cadre de sa thèse, du contrôle de populations d'insectes par technique de l'insecte stérile. Enfin la combinaison de jeux de données que j'ai présentée dans le Chapitre 3 se prolongent d'une part dans la thèse en cours d'Emma Thulliez, dans le cadre de l'estimation de cartes de pollution de l'air à partir de mesures de différentes qualité, et d'autre part sur la combinaison de données de comptages et de données de séquençages pour estimer les paramètres démographiques d'une population, qui sera étudiée par Lucas Rey dans le cadre de son post-doctorat. Ces sujets sont très différents les uns des autres mais peuvent se rejoindre. En particulier la technique de l'insecte stérile pourra avantageusement être étudiée par combinaison de données démographiques et génétiques.

Bibliographie

D. Abu Awad et C. Coron : Effects of demographic stochasticity and life-history strategies on times and probabilities to fixation. *Heredity*, 121(4):374–386, 2018.

Aneil Agrawal : Sexual selection and the maintenance of sexual reproduction. *Nature*, 411:692–5, 07 2001.

K B Athreya et P E Ney : *Branching processes*. Springer-Verlag Berlin, Mineola, NY, 1972. ISBN 0-486-43474-5. Reprint of the 1972 original [Springer, New York ; MR0373040].

Jonathan B.L. Bard : Modelling speciation : Problems and implications. *In Silico Biology*, 15 (1-2):23–42, 2023.

Arnaud Becheler : *Environmental demogenetic model*. Thèse de doctorat, Université Paris-Saclay, 2018. URL http://www.theses.fr/2018SACLS145. Thèse de doctorat dirigée par Stéphane Dupas, Biologie, Université Paris-Saclay (ComUE) 2018.

M.S. Boyce et L.L. McDonald : Relating populations to habitats using resource selection functions. *Trends in Ecology & Evolution*, 14:268–272, 1999.

Nicolas Champagnat : *Etude mathématique de modèles stochastiques d'évolution issus de la théorie écologique des dynamiques adaptatives*. Thèse de doctorat, Université Paris Nanterre, 2004. URL http://www.theses.fr/2004PA100138. Thèse de doctorat dirigée par Sylvie Méléard, Mathématiques, Paris 10 2004.

Nicolas Champagnat : A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic Process. Appl.*, 116(8):1127–1160, 2006. ISSN 0304-4149. URL http://dx.doi.org/10.1016/j.spa.2006.01.004.

Nicolas Champagnat, Régis Ferrière et Sylvie Méléard : Unifying evolutionary dynamics : From individual stochastic processes to macroscopic models. *Theor. Popul. Biol.*, 69:297–321, 2006.

Nicolas Champagnat et Sylvie Méléard : Invasion and adaptive evolution for individual-based spatially structured populations. *Journal of Mathematical Biology*, 55(2):147–188, 2007.

Joseph T. Chang : Recent common ancestors of all present-day individuals. *Advances in Applied Probability*, 31(4):1002–1026, 1999.

C Chicone : *Ordinary Differential Equations with Applications*. Numéro 34 *in* Texts in Applied Mathematics. Springer-Verlag New York, 2006.

C Coron : Slow-fast stochastic diffusion dynamics and quasi-stationarity for diploid populations with varying size. *Journal of Mathematical Biology*, pages 1–32, 2015.

C. Coron, C. Calenge, C. Giraud et R. Julliard : Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. *Ecological and environmental statistics*, 25:71–93, 2018a.

C. Coron, M. Costa, F. Laroche, H. Leman et C. Smadi : Emergence of homogamy in a two-loci stochastic population model. *ALEA, Lat. Am. J. Probab. Math. Stat.*, 18(1):469–508, 2021.

C. Coron, M. Costa, H. Leman, V. Llaurens et C. Smadi : Origin and persistence of polymorphism in loci targeted by disassortative preference : a general model. *J Math Biol.*, 86 (1), 2022.

C. Coron et Y. Le Jan : Pedigree in the biparental moran model. *J. Math. Biol.*, 84, 2022.

C. Coron et Y. Le Jan : Genetic contribution of an advantaged mutant in the biparental moran model. *Ukr Math J*, 75:1666–1672, 2024a.

C. Coron et Y. Le Jan : Genetic contribution of an advantaged mutant in the biparental moran model - finite selection. *ArXiv*, 2024b.

C. Coron, S. Méléard et D. Villemonais : Impact of demography on extinction/fixation events. *J. Math. Biol.*, 78:548–577, 2019.

Camille Coron, Manon Costa, Hélène Leman et Charline Smadi : A stochastic model for speciation by mating preferences. *Journal of mathematical biology*, 76(6):1421–1463, 2018b.

Camille Coron et Yves Le Jan : Genetic contribution of an advantaged mutant in the biparental Moran model - finite selection. URL https://hal.science/hal-04570119. working paper or preprint, mai 2024c.

Élisa Couvert, François Bienvenu, Jean-Jil Duchamps, Adélie Erard, Verónica Miró Pina, Emmanuel Schertzer et Amaury Lambert : Opening the species box : What parsimonious microscopic models of speciation have to say about macroevolution. *bioRxiv*, 2024. URL https://www.biorxiv.org/content/early/2024/10/03/2023.11.09.564915.

Charles Darwin : *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

Richard Dawkins : *The selfish gene*. Oxford University Press, New York, 1976. ISBN 019857519X.

Bernard Derrida, Susanna C. Manrubia et Damian H. Zanette : On the genealogy of a population of biparental individuals. *Journal of Theoretical Biology*, 203(3):303 – 315, 2000.

S N Ethier et T G Kurtz : Markov processes : Characterization and convergence, 1986, 1986.

Thomas Flatt et Andreas Heyland : *Mechanisms of Life History Evolution : The Genetics and Physiology of Life History Traits and Trade-Offs*. Oxford University Press, 05 2011. ISBN 9780199568765.

M.I. Freidlin, J. Szücs et A.D. Wentzell : *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer, 1984.

Hans-Otto Georgii et Ellen Baake : Supercritical multitype branching processes : the ancestral types of typical individuals. *Advances in Applied Probability*, 35(4):1090–1110, 2003.

Christophe Giraud, Clément Calenge, Camille Coron et Romain Julliard : Capitalizing on opportunistic data for monitoring species relative abundances. *Biometrics*, 72(2):649–58, 2015.

Hans-Rolf Gregorius : A two-locus model of speciation. *Journal of theoretical Biology*, 154 (3):391–398, 1992.

M Herrero : Male and female synchrony and the regulation of mating in flowering plants. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 358:1019–1024, 2003.

O P Höner, B Wachter, M L East, W J Streich, K Wilhelm, T Burke et H Hofer : Female mate-choice drives the evolution of male-biased dispersal in a social mammal. *Nature*, 448:797–802, 2007.

Frédéric Jiguet, Vincent Devictor, Romain Julliard et Denis Couvet : French citizens monitoring ordinary birds provide tools for conservation and ecological sciences. *Acta Oecologica*, 44:58 – 66, 2012.

Motoo Kimura et Tomoko Ohta : *CHAPTER EIGHT. Breeding Structure of Populations*, pages 117–140. Princeton University Press, 2020. URL https://doi.org/10.12987/9780691210094-009.

Amaury Lambert, Verónica Miró Pina et Emmanuel Schertzer : Chromosome painting. *arXiv :1807.09116*, 2018.

Subhash R. Lele, Evelyn H. Merrill, Jonah Keim et Mark S. Boyce : Selection, use, choice and occupancy : clarifying concepts in resource selection studies. *Journal of Animal Ecology*, 82:1183–1191, 2013.

RC Lewontin, LR Ginzburg et SD Tuljapurkar : Heterosis as an explanation for large amounts of genic polymorphism. *Genetics*, 88(1):149–169, 1978.

Martin Linder : Common ancestors in a generalized Moran model. *U.U.D.M. Reports*, 2009.

Ludovic Maisonneuve, Thomas Beneteau, Mathieu Joron, Charline Smadi et Violaine Llaurens : When do opposites attract ? a model uncovering the evolution of disassortative mating. *The American Naturalist*, 198(5):000–000, 2021.

Lucas Marie-Orleach, Sylvain Glémin, Marie K. Brandrud, Anne K. Brysting, Abel Gizaw, A. Lovisa S. Gustafsson, Loren H. Rieseberg, Christian Brochmann et Siri Birkeland : How does selfing affect the pace and process of speciation ? *Cold Spring Harbor Perspectives in Biology*, 2024. URL http://cshperspectives.cshlp.org/content/early/2024/03/19/cshperspect.a041426.abstract.

D K McLain et R D Boromisa : Male choice, fighting ability, assortative mating and the intensity of sexual selection in the milkweed longhorn beetle, tetraopes tetraophthalmus (coleoptera, cerambycidae). *Behavioral Ecology and Sociobiology*, 20(4):239–246, 1987.

GA McVean et NJ Cardin : Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–93, 2005.

J.A.J. Metz, S.A.H. Geritz, G. Meszéna, F.J.A. Jacobs et J.S. Van Heerwaarden : Adaptive dynamics : A geometrical study of the consequences of nearly faithfull reproduction. *In Stochastic and spatial structures of dynamical systems*, pages 183–231. S.J. van Strien and S.M. Verduyn-Lunel (eds.), North Holland, Elsevier, 1996.

L K M'Gonigle, R Mazzucco, S P Otto et U Dieckmann : Sexual selection enables long-term coexistence despite ecological equivalence. *Nature*, 484(7395):506–509, 2012.

Maximillian Newman, John Wakeley et Wai-Tong Louis Fan : Conditional gene genealogies given the population pedigree for a diploid moran model with selfing, 2024. URL https://arxiv.org/abs/2411.13048.

R J H Payne et D C Krakauer : Sexual selection, space, and speciation. *Evolution*, 51(1):1–9, 1997.

Peter Pfaffelhuber et Anton Wakolbinger : A diploid population model for copy number variation of genetic elements. *Electronic Journal of Probability*, 28(none):1 – 15, 2023. URL https://doi.org/10.1214/23-EJP934.

Martyn Plummer : Jags : A program for analysis of bayesian graphical models using gibbs sampling, 2003.

Martyn Plummer : *rjags : Bayesian graphical models using MCMC*, 2014. URL http://CRAN.R-project.org/package=rjags. R package version 3-13.

R Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

V Savolainen, M C Anstett, C Lexer, I Hutton, J J Clarkson, M V Norup, M P Powell, D Springate, N Salamin et W J Baker : Sympatric speciation in palms on an oceanic island. *Nature*, 441:210–213, 2006.

P L Schwagmeyer : Scramble-competition polygyny in an asocial mammal : Male mobility and mating success. *The American Naturalist*, 131:885–892, 1988.

Kerry L. Shaw, Christopher R. Cooney, Tamra C. Mendelson, Michael G. Ritchie, Natalie S. Roberts et Leeban H. Yusuf : How important is sexual isolation to speciation ? *Cold Spring Harbor Perspectives in Biology*, 16(4), 2024. URL http://cshperspectives.cshlp.org/content/16/4/a041427.abstract.

Bruno O. Shubert : A flow-graph formula for the stationary distribution of a markov chain. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-5(5):565–566, 1975.

Lionel Soulhac, Chi Vuong Nguyen, P Volta et P Salizzoni : The model sirane for atmospheric urban pollutant dispersion. part iii : validation against no2 yearly concentration measurements in a large urban agglomeration. *Atmospheric environment*, 167:377–388, 2017.

Thorsten Stefan, Louise Matthews, Joaquin M Prada, Colette Mair, Richard Reeve et Michael J Stear : Divergent allele advantage provides a quantitative model for maintaining alleles with a wide range of intrinsic merits. *Genetics*, 212(2):553–564, 2019.

université
PARIS-SACLAY

Titre : Etude probabiliste de la biodiversité et de la génétique des populations à reproduction sexuée

Mots clés : Modèles probabilistes ; Biodiversité et diversité génétique ; Convergence des suites de processus stochastiques et comportement en temps long ; Sciences participatives ; Fusion de données.

Résumé : Ce manuscrit présente une partie de mes travaux de recherche qui se situent en probabilités pour la biologie. Il est constitué de trois parties, qui portent sur des questions biologiques distinctes, abordées avec des modèles mathématiques différents, et des groupes de collaborateurs différents. Dans la première partie j'étudie la composition génétique d'une population à reproduction biparentale. Plus précisément, avec Yves Le Jan nous avons étudié le comportement de la proportion de génome transmise en temps long par un ancêtre ou un groupe d'ancêtres, dans un modèle de Moran biparental avec et sans sélection. Dans la deuxième partie j'étudie le rôle des préférences d'appariement dans l'évolution génétique des populations. Plus précisément, avec Manon Costa, Hélène Leman et Charline Smadi nous avons étudié, à l'aide de modèles probabilistes individu-centrés, (i) les conditions d'émergence de l'homogamie, (ii) le rôle de l'homogamie dans la spéciation, et (iii) le niveau de diversité génétique généré par l'hétérogamie. Dans la troisième partie j'étudie comment des données d'observation de la biodiversité réalisées par des citoyens et collectées via des programmes de sciences participatives peuvent améliorer l'évaluation et le suivi de la biodiversité. Plus précisément, avec Clément Calenge, Christophe Giraud, et Romain Julliard nous avons montré que la combinaison, au travers d'un modèle probabiliste simple, de ces données avec des données récoltées par des professionnels permet l'élaboration de cartes d'abondances d'espèces plus précises que celles obtenues par le seul jeu de données professionnel.

Title : Probabilistic study of biodiversity and population genetics of sexually reproducing populations

Keywords : Probabilistic models ; Biodiversity and genetic diversity ; stochastic processes : scalings, convergence and asymptotic behavior ; citizen sciences ; data fusion

Abstract : This manuscript presents part of my research work, that lies in Probability for Biology. It is divided in three parts, which deal with distinct biological questions, studied using different mathematical models, and with different groups of collaborators. In the first part I study the genetic composition of a population with sexual reproduction. More specifically, with Yves Le Jan we have studied the asymptotic proportion of the genome transmitted by an ancestor or a group of ancestors, in a biparental Moran model with and without selection. In the second part, I study the role of mating preferences in the genetic Évolution of populations. More specifically, with Manon Costa, Hélène Leman and Charline Smadi, using individual-based probabilistic models, we studied first the conditions for the emergence of homogamy, second the role of homogamy in speciation, and third the level of genetic diversity generated by heterogamy. In the third part, I study how biodiversity observation data collected by citizens via citizen science programs can improve the assessment and monitoring of biodiversity. More specifically, with Clément Calenge, Christophe Giraud and Romain Julliard, we proved that combining these data with data collected by professionals, through a simple probabilistic model, makes it possible to produce more accurate species abundance maps than those obtained from the professional dataset alone.