Probabilistic population genetics models for expanding populations

Apolline Louvet

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX026

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Modèles probabilistes de génétique des populations pour les populations en expansion

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 7 juin 2022, par

APOLLINE LOUVET

Composition du Jury :

Nicolas Champagnat Directeur de recherche, IECL, Université de Lorraine	Rapporteur
Jochen Blath Professor, TU Berlin	Rapporteur
Sylvie Méléard Professeure, CMAP, Ecole Polytechnique	Examinatrice
Guillaume Achaz Professeur, Université Paris Cité	Président du jury
Fabien Laroche Ingénieur des Ponts, des Eaux et des Forêts, INRAE	Examinateur
Amandine Véber Directrice de recherche, MAP5, Université Paris Cité	Directrice de thèse
Nathalie Machon Professeure, CESCO, MNHN	Co-directrice de thèse

Résumé

Cette thèse porte sur la construction et l'étude de modèles probabilistes de génétique des populations pour les populations en expansion. Nous construisons ces modèles à partir d'un concept issu de la théorie des systèmes de particules en interaction, qui consiste à représenter les sites vides comme occupés par des particules d'un type spécifique. Ces "individus fantômes" nous permettent de maintenir artificiellement un nombre d'individus constant, et de construire des processus duaux encodant les généalogies. Dans nos modèles, les individus fantômes peuvent aussi se reproduire, modélisant ainsi les fluctuations stochastiques du nombre d'individus, mais avec un très fort désavantage sélectif face aux individus "réels". Nous appliquons d'abord le concept d'individus fantômes à un processus à valeurs mesure qui décrit la dynamique de reproduction d'une population vivant dans un espace continu. Nous construisons la limite de ce processus lorsque la "sélection" contre les individus fantômes devient infiniment forte. Le processus limite semble être un équivalent du modèle d'Eden en espace continu. Nous étudions la dynamique d'expansion des individus réels dans le processus limite, et montrons que la croissance de la région qu'ils occupent est linéaire en temps. Nous nous intéressons ensuite à une variante du modèle de Wright-Fisher structuré spatialement, incluant une banque de graines et des extinctions locales fréquentes. Ceci est motivé par une question d'intérêt en écologie : comprendre la dynamique des plantes dans les pieds d'arbres d'alignement en ville. Dans une étude préliminaire sur un jeu de données réelles, nous montrons qu'il est nécessaire de prendre en compte la présence potentielle d'une banque de graines pour répondre à cette question. Nous utilisons notre variante du modèle de Wright-Fisher pour montrer l'existence d'une probabilité critique d'extinction de patch dépendant des paramètres de banque de graines au delà de laquelle une expansion de population n'est pas possible. Nous étudions la limite de ce processus dans un régime de sélection forte, et montrons qu'il converge vers un modèle de présence/absence. Ce modèle limite appartient à une famille de modèles très utilisés en écologie des métapopulations.

Abstract

This thesis focuses on the construction and study of stochastic population genetics models for expanding populations. We build different models using a concept from the theory of interacting particles systems, where empty sites are represented as occupied by particles with a specific type. These "ghost individuals" allow us to artificially keep population sizes constant, and build dual processes encoding genealogies. In our models, ghost individuals can reproduce as well in order to account for stochastic fluctuations in population sizes, but with a very strong selective disadvantage against "real" individuals. We first apply the concept of ghost individuals to a measure-valued process describing the reproduction dynamics of a population living in a continuous space. We construct the limit of the process when "selection" against ghost individuals becomes infinitely strong. The limiting process seems to be a space continuous equivalent of the Eden growth model. We study the expansion dynamics of real individuals in the limiting process, and show that the growth of the region they occupy is linear in time. We then focus on a variant of the spatially-structured Wright-Fisher model with a seed bank component and featuring frequent local extinction events. This was motivated by a question of ecological interest: understanding plant dynamics in urban tree bases. In a preliminary study on a real dataset, we show that it is necessary to account for the potential presence of a seed bank in order to answer this question. We use our variant of the Wright-Fisher model to show the existence of a critical patch extinction probability depending on

seed bank parameters, above which a population expansion is not possible. We study the limit of the process under a strong selection regime, and show that it corresponds to an occupancy-based model. This limiting process belongs to a family of models widely used in metapopulation ecology.

Remerciements

A la sortie du lycée, même si la biologie et l'écologie m'intéressaient beaucoup, je préférais les mathématiques, et je n'avais aucune idée du fait qu'il était possible de faire de la recherche à l'interface entre ces différents domaines. Je pensais qu'il me fallait choisir, et je me suis donc orientée vers les mathématiques avec pour objectif de faire de la recherche dans un domaine tel que la géométrie algébrique (bien entendu, je n'avais aucune idée de ce que c'était, mais le nom sonnait bien). Neuf ans plus tard, me voici finalement sur le point de soutenir une thèse conjuguant à la fois mathématiques et écologie, et c'est en grande partie grâce à Amandine.

En effet, mon premier contact avec les mathématiques appliquées à la biologie a eu lieu dans le cadre d'un groupe de travail organisé par Amandine. Rassemblant élèves des départements de mathématiques et de biologie, il a représenté pour moi une véritable porte d'entre sur le monde de la recherche en maths-bio, aussi bien du point de vue des thématiques (des micronageurs à l'épidémiologie, en passant par la génétique des populations) que de celui de la dynamique d'interaction avec biologistes et écologues. Merci Amandine de m'avoir ensuite acceptée en stage, puis en thèse, durant lesquels tu m'auras transmis ta passion des interactions entre mathématiques et biologie. Merci pour tes conseils et encouragements dans la rédaction d'articles, et pour m'avoir donné des pistes de recherche toujours intéressantes (bon peut-être pas les intégrales infernales sur les ellipses :-)) tout en me laissant aussi identifier et explorer d'autres questions par moi-même. J'espère un jour réussir à percer le secret de comment tu arrives à t'impliquer dans autant de commissions, comités scientifiques et projets tout en restant toujours aussi disponible pour tes doctorants ! En attendant, j'essaierai de suivre ton exemple et de moi aussi, à mon échelle, contribuer à la vie du laboratoire et de la communauté scientifique.

Merci à Nathalie d'avoir accepté d'accueillir au CESCO une matheuse en mission d'infiltration visant à comprendre comment échanger avec ces individus mystérieux que sont les écologues. Ce qui devait au départ n'être qu'un stage de six mois s'est finalement transformé en thèse, et je te suis extrêmement reconnaissante de tout ce qu'il m'a apporté. Merci pour m'avoir donné l'opportunité de me confronter aux différents aspects plus pratique de la recherche en maths-bio, tels que l'analyse ou la collecte de données. Et surtout, merci pour ta patience, tes encouragements, pour reconnaître malgré tout un certain esthétisme aux équations d'une page, et pour m'avoir poussée à changer le nom des "individus inexistants".

I am very grateful to Jochen Blath and Nicolas Champagnat for having accepted to review my thesis, and for your careful reading of the manuscript. Je remercie également Guillaume Achaz, Fabien Laroche et Sylvie Méléard pour avoir accepté de faire partie de mon jury de thèse.

Pour diverses raisons, j'ai été amenée à passer ma thèse dans quatre endroits différents : le CMAP, le MAP5, le CESCO, et l'annexe bordelais du CMAP5 (plus communément désigné sous le nom de "domicile de mes parents"). Merci au CESCO pour m'avoir accueilli en stage puis plus ponctuellement pendant ma thèse, pour me laisser infiltrer les Incroyables Journées du Rocheton, et pour avoir enrichi mon répertoire d'anecdotes à sortir à table de nombreux *fun facts* sur les chauves souris. J'espère avoir l'occasion de continuer à croiser certains d'entre vous lors des journées de la chaire MMB ! Merci en particulier à Jean-Baptiste Mihoub et Alexandre Robert pour m'avoir accompagnée tout au long de ma première confrontation au monde des données réels, et de ma première expérience d'écriture d'article. Merci pour vos conseils, vos encouragements, vos retours toujours constructifs, et pour nos échanges pendant le premier confinement. Merci aussi aux autres doctorantes et doctorants de l'équipe : Mona, Laura Chloé, Eduardo, Tanguy et Hortense. J'ai vraiment apprécié échanger avec vous, et votre intérêt pour mes thématiques de recherche, même

si elles sont assez éloignées des vôtres. Bonne chance à Hortense pour ta fin de thèse, à Eduardo et Tanguy pour votre soutenance prochaine, et à Chloé pour la suite !

Merci au MAP5 pour m'avoir accueillie aussi chaleureusement lors de mon arrivée dans les bagages d'Amandine il y a deux ans, et pour les discussions sur la terrasse ensoleillée l'été ou près du chauffage l'hiver. Même si la terrasse est à elle seule une bonne raison de venir, c'est surtout pour la convivialité légendaire du laboratoire que j'affronte le RER B ! Merci en particulier à l'équipe de probabilités, et à Marie-Hélène pour ton aide pour les départs en conférence et l'administratif associé. Merci à l'ensemble des doctorantes et doctorants du labo pour les repas, le père noël secret, le laser-game, ainsi que nos aventures à Guidel, Besançon, ou dans les contrées reculées du plateau de Saclay. Merci en particulier à mes frères et sœurs de thèse Anne (bonne chance pour ta soutenance !), Laurent et Emilie (merci à Léonard et toi pour votre aide pour la préparation des entretiens de post-doc !), à mes camarades du bureau 750, et à Zoé pour ta bonne humeur et ta gentillesse.

Merci à mes parents pour m'avoir laissé ouvrir un annexe du CMAP5 à Bordeaux durant les deuxième et troisième confinements. Grâce à vous, cette période compliquée s'est révélée être l'un des moments les plus productifs de ma thèse. Merci à Falbala et Minouchette pour m'avoir laissé m'installer dans leur chambre pour travailler. Désolée pour les bruits de clavier et les réunions Zoom qui ont dérangé vos siestes !

Et surtout, *last but not least*, merci au CMAP ! Cela fait maintenant cinq ans que je suis venue pour la première fois au CMAP dans le cadre d'un stage, et je pense que tout me manquera : le cadre, le Magnan, l'ambiance, l'équipe administrative et son efficacité (merci en particulier à Nasséra et Alex !), et peut-être même le 91.06. Merci au groupe PEIPS pour votre convivialité et votre bonne humeur, les séances du groupe de travail et les goûters associés, et tous nos échanges. Merci à Carl Graham pour nos discussions autour des modèles d'expansion, des techniques de preuves associées, et pour m'avoir aidé à connecter le modèle de génétique des populations que j'étudiais à ceux utilisées dans d'autres branches des probabilités. Merci à Igor Kortchemski pour ce qui aura été mon premier semestre de monitorat, ton implication dans le cours, les réunions du lundi midi au Magnan, et pour m'avoir poussée à m'investir comme représentante des moniteurs du département. Donner les TDs de ton cours aura été pour moi une excellente introduction au monde de l'enseignement. Merci aussi à François Alouges, à Thibaut Mastrolia, à ma camarade de fronde anti-DMs Milica, ainsi qu'à Benoît, Claire, Shanqing, Kang et Ariane.

Merci aux doctorantes et doctorants du CMAP, et en particulier à mes collègues passés et présents du bureau 20.03 (en toute objectivité le meilleur bureau de doctorants du CMAP). Merci aussi à mes camarades d'organisation du CJC-MA: Constantin, Thomas, Claire, Dominik, Solange, Clément, Baptiste, Louis, Josué, Pierre, Corentin et Guillaume. Cela aura été une super expérience, très riche d'enseignements, et contrairement à nos prévisions les plus pessimistes, nous ne nous sommes pas tous entre-tués avant la fin du congrès :-) Avec le recul, se lancer dans l'organisation d'un nouveau congrès, en pleine pandémie mondiale et sans vraiment avoir d'idée de ce que cela impliquait était un sacré pari, mais je pense que nous l'avons relevé haut la main. Cela n'aurait pas été possible sans le soutien de Matthieu Aussal, qui nous a encouragé à nous lancer dans ce projet et nous a aidé tout au long de l'aventure. Merci aussi à la SMAI, au CMAP et à l'Ecole Polytechnique pour avoir soutenu l'événement.

Et comme il y a aussi une vie en dehors de la thèse, merci à tous ceux avec qui j'ai partagé soirées sushi, tables de JdR ou répétitions, et qui m'ont soutenu dans les moments difficiles. Merci au bureau de l'école de musique Pierre Paubon, à mes camarades de répétition, à Julia pour tout ce que tu m'as apporté musicalement ces trois dernières années (bon courage pour ta thèse !), et à Elisa pour nos répétitions/thé. Thanks Cristobal, Marcella, Dominik, Hanieh and Shri for all the

hours we spent killing zombies, adopting cultists and complaining about the Welcomers not being that welcoming (though for this last point, that may have been me only). Merci à Marc et Laure pour tous nos repas et expéditions au sushi à volonté (vous êtes les bienvenus à Bath, mais je ne promets pas la présence du barbecue électrique !), et merci à Claude pour les soirées crêpe. Merci à Michel pour les découvertes culinaires, les randonnées, les meringues-crâne, et pour m'avoir permis d'infiltrer le club de JdR de l'ENS Saclay. Merci à celles et ceux que j'y ai croisé en murder, avec qui j'ai (attention spoilers) colmaté des fuites avec de la purée, échangé des ogives nucléaires, ou fomenté des complots pour renverser un mage millénaire. Merci à Ikhlas et Inès, vous êtes vous aussi les bienvenues à Bath ! Thanks to Sarjick and Himalay for always having been here for me since we met in CMI, and for helping me on my path to write English with a less broken grammar :-)

Enfin, merci à ma famille pour m'avoir toujours soutenue. Cette thèse vous est dédiée.

Summary

Part I

Introduction

Chapter 1

Introduction (en français)

1.1 Motivations biologiques

Les expansions de population sont un phénomène courant qui se produit à toutes les échelles de temps, d'espace et d'organisation du vivant. Les exemples en sont variés : espèces invasives, tumeurs cancéreuses, mais aussi certaines espèces d'arbres en Europe [SNS08; SS07], sans oublier bien sûr l'exemple (donné par l'actualité récente) des expansions de variants successifs d'un virus. De plus, beaucoup d'espèces dont l'aire de répartition peut actuellement être considérée comme stable ont en fait été en expansion dans le passé, suite à une modification du climat, à l'introduction de l'espèce dans un nouveau milieu, ou encore à l'apparition de nouvelles mutations. Par exemple, la plante *Veronica persica* (voir Figure 1.1), aux petites fleurs bleues et blanches, est aujourd'hui présente dans toute la France métropolitaine [MO22], et est souvent présente dans les parcs et jardins. Pourtant, il s'agit en fait une espèce originaire du sud-ouest de l'Asie, qui a été introduite un peu partout dans le monde et qui est considérée comme invasive dans de nombreux pays. Sur une échelle de temps plus longue, la plupart des espèces animales et végétales que nous observons tous les jours en Europe ont été en expansion à la fin de la dernière période glaciaire, et ont (re)colonisé tout ou partie de l'Europe [Wil+93]. Plusieurs espèces d'arbres [SNS08; SS07], d'herbes [Van+07], de reptiles ou d'amphibiens [Ara+08; APR05] n'ont d'ailleurs toujours pas fini, et sont encore en expansion.

Figure 1.1: *Veronica persica*

Les expansions de population laissent une empreinte durable sur les patterns de diversité génétique dans la population, qui reste encore visible même une fois l'expansion terminée. En effet, deux individus sont d'autant plus proches génétiquement qu'ils ont un ou des ancêtres communs récent. Dans le cas d'individus haploïdes (i.e, de façon simplifiée, d'individus ayant un seul parent), ceci est encore plus marqué, et les seules différences entre deux individus ayant un ancêtre commun ne peuvent qu'être liées aux mutations s'étant produites depuis cet ancêtre. La diversité génétique observée est donc directement liée aux relations de parenté entre les individus, et par extension à

leurs généalogies. Or, une expansion de population conduit à des généalogies différentes de celles associées à des populations à l'équilibre. Ainsi, lorsque nous remontons dans le temps et reconstruisons les généalogies d'un échantillon d'individus, les ancêtres sont de plus en plus proches de la zone d'où est partie l'expansion. De plus, les individus proches du front, où les densités de population sont plus faibles pour une même quantité de ressources, sont moins en compétition et peuvent donc généralement se reproduire plus.

D'autres événements, tels que des événements d'extinction, vont eux aussi affecter les généalogies, et c'est ainsi que la diversité génétique observée au temps présent contient des informations sur le passé de la population. Exploiter cette information permet par exemple de comprendre comment les espèces se sont adaptées aux changements climatiques passés [Hew00], ou de reconstituer l'histoire évolutive humaine [Tem02]. Dans un tout autre champ d'application, les patterns de diversité génétique dans une tumeur cancéreuse peuvent permettre d'en apprendre plus sur les événements s'étant produits au début de sa phase de croissance, lorsqu'elle était trop petite pour être détectée et analysée. L'un des axes de recherche en génétique des populations consiste ainsi à développer des outils permettant d'extraire de l'information de la diversité génétique observée, afin de reconstituer le passé d'une population. Plus loin dans l'introduction (Sections 1.1.1 et 1.1.2), j'expliquerai pourquoi développer ces techniques d'inférence est plus difficile dans le cas de populations en expansion, et quels sont les apports de ma thèse pour répondre à cette problématique.

Les facteurs déclencheurs d'une expansion de population sont variés : généralement, il s'agit de l'introduction d'une espèce dans un nouveau milieu qui lui est favorable, d'une modification de l'environnement ou du climat, ou de l'apparition d'une nouvelle mutation rendant les individus qui la portent plus compétitifs. Pour autant, comprendre comment déclencher ou empêcher une expansion de population en pratique n'est pas toujours facile, en témoignent l'échec de certains programmes de réintroduction d'espèces menacées ou (à l'inverse) d'éradication d'espèces invasives. Pour illustrer cela, considérons l'exemple d'une mutation délétère (c'est-à-dire conduisant à une diminution du nombre de descendants). Dans une population à l'équilibre, les individus qui portent cette mutation se reproduisent moins vite, et sont en compétition avec les autres. La sous-population d'individus portant la mutation délétère est vouée à s'éteindre à court ou moyen terme, même en l'absence de mesures mises en place pour contrôler ou empêcher son expansion. En revanche, ce n'est pas toujours le cas dans une population en expansion. En effet, la mutation peut subsister sur des échelles de temps plus ou moins longues en "surfant" sur le front, là où les densités de population sont plus faibles et donc la compétition moins forte (mais la dérive génétique plus forte) [FE20; Pei+13; Tra+07]. De plus, si la mutation augmente la capacité de dispersion, il est possible sous certaines conditions que la mutation envahisse complètement le front, malgré son désavantage sélectif ("survival of the fastest", [CP16; Def+19]). A long terme, la sous-population correspondante finira par s'éteindre, mais pas avant que l'expansion ne soit terminée et que la totalité de l'aire de répartition possible ne soit occupée. De plus, l'expansion de la sous-population d'individus ne *portant pas* la mutation s'en trouve freinée. Le phénomène de surf de mutations délétères sur le front a été observé en laboratoire [Dit+13; Def+19], mais aussi dans la nature avec plusieurs espèces invasives, comme le crapaud buffle en Australie [Hud+15] (voir l'introduction de [Def+19] pour d'autres exemples).

L'exemple de mutations délétères dans une population en expansion, qui est aussi généralisable au cas d'une espèce ayant un désavantage sélectif face à une autre espèce avec laquelle elle est en compétition, illustre bien les phénomènes a priori contre-intuitifs qui peuvent se produire lors d'une expansion de population, et qui empêchent de comprendre ce qui favorise ou bloque l'expansion d'une (sous) population. Pourtant, répondre à cette question est un enjeu majeur pour lutter contre les espèces invasives, qui représentent une menace pour la biodiversité [BCB16; BBR19], les écosystèmes [Kum+15; WCV16], l'économie [Bra+16; Dia+21], l'agriculture [Pai+16] et/ou la santé [She+11]. A une toute autre échelle biologique, les tumeurs cancéreuses sont elles aussi des

populations en expansion, et comprendre comment bloquer leur croissance pourrait permettre le développement de nouvelles thérapies. De plus, comprendre ce qui *empêche* une expansion de population pourrait présenter des applications pour des programmes de réintroduction ou de protection d'espèces, ou pour augmenter la connectivité dans des environnements fragmentés. Durant ma thèse, je me suis focalisée sur deux questions biologiques d'intérêt liées aux populations en expansion, que je vais maintenant présenter plus en détail.

1.1.1 Diversité génétique au front d'une population en expansion

Résultats expérimentaux

La première question à laquelle je me suis plus particulièrement intéressée est motivée par une expérience de Oskar Hallatschek et collaborateurs [Hal+07; HN10]. Cette expérience consiste à mettre en croissance des souches fluorescentes de bactéries (*E. coli*) ou de levures (*S. cerevisiae*) dans des boîtes de Pétri, en les plaçant au centre de la boîte après les avoir mélangées. La couleur de la fluorescence est codée par un seul gène, et est la seule différence entre les deux souches. De plus, elle est neutre sélectivement, et n'influence pas le comportement des bactéries. Autrement dit, elle peut être vue comme un simple marqueur, permettant de visualiser la répartition des types de bactéries.

Les résultats obtenus sont les suivants. Si les deux souches fluorescentes sont toujours bien mélangées dans la zone initialement occupée par les bactéries ou les levures, la situation est très différente au niveau du front d'expansion. Celui-ci est en effet composé de secteurs dans lesquels tous les individus émettent une fluorescence de la même couleur. Les secteurs sont de plus relativement larges et aux frontières très nettes. La forme et le nombre de secteurs dépend de l'espèce considérée, ainsi que de la forme initiale de la colonie (ligne ou disque). Voir [Hal+07] pour une illustration du phénomène.

L'apparition de ces secteurs est due à la combinaison des deux facteurs : de plus faibles densités de population au niveau du front, conduisant à une forte dérive génétique (i.e, à de fortes fluctuations stochastiques des fréquences des différents types), et un taux de reproduction plus élevé dans cette zone (du fait d'une compétition moindre) qui amplifie l'effet de la dérive génétique. Concrètement, dans le cas de la dimension 1, du fait de la dérive génétique, l'un des deux types peut s'éteindre au front. Il est a priori toujours présent plus en arrière, là où les densités de population sont plus fortes. Mais pour "rattraper" le front d'expansion, il est maintenant en compétition avec un nombre plus élevé d'individus de l'autre type : ses voisins et les individus formant le front. Le front "actualisé" ne sera alors composé avec grande probabilité que des individus de l'autre type, qui pourront là encore contribuer plus facilement à la prochaine avancée du front. Voir la Figure 1.2 pour une illustration du phénomène, qui peut être considéré comme des effets de fondation successifs.

Ainsi, en dimension 1, l'expansion de population conduit à une diminution de la diversité génétique dans la direction de l'expansion, par un mécanisme de nature principalement stochastique, qui ne peut être approché par un modèle déterministe que sous certaines conditions [HN08]. Le gradient de diversité génétique qui en résulte a principalement été étudié via des simulations [EFP09; HN08], et plus récemment analytiquement [DF16]. Cette décroissance de la diversité génétique dans la direction d'expansion peut être observée dans des populations réelles (voir par exemple [Hew96; Mac+96; Ros+02; Ram+05]), et peut être utilisée pour reconstituer des expansions passées (voir par exemple [Hew00; Tem02]). En dimension 2 ou plus, la même chose se produit dans chaque direction d'expansion. Cependant, cette fois-ci, le type fixé dans des directions différentes n'est pas forcément le même. L'émergence des secteurs est alors due aux corrélations entre les types génétiques se fixant dans des directions différentes mais proches.

En plus de l'expérience présentée plus haut, ces secteurs caractéristiques des populations en expansion ont aussi été observés dans des populations réelles de tortues [Gra+13] et dans des

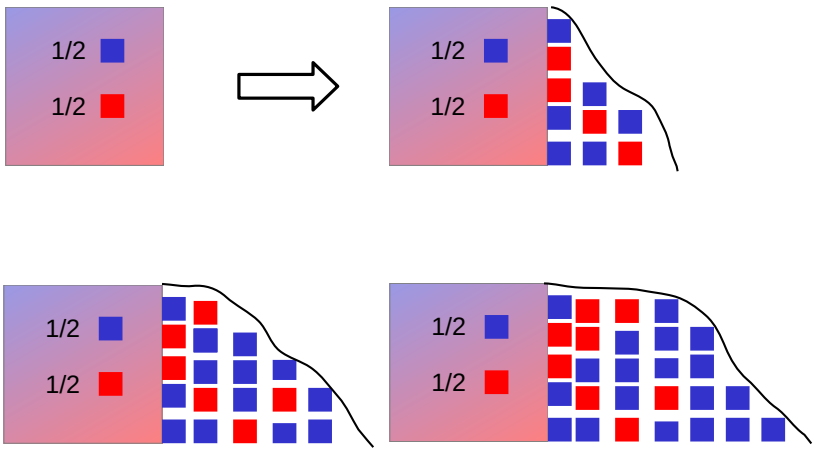


Figure 1.2: Décroissance de la diversité génétique dans le front d'une population en expansion. Un grand nombre d'individus sont présents dans la zone de laquelle part l'expansion. Les deux types possibles (bleu ou rouge) sont neutres sélectivement et initialement présents dans les mêmes proportions.

tumeurs cancéreuses [Sot+15]. Pour autant, la présence de zones dans lesquelles tous les individus ont le même type n'est pas forcément le marqueur d'une expansion passée ou en cours, et peut simplement être due à de la sélection naturelle variable dans l'espace, en faveur de l'un ou l'autre des deux types. Ceci est un obstacle à l'utilisation de données génétiques pour reconstituer des expansions passées, mais aussi pour identifier des mutations avantagées sélectivement [Cur+06; Sch+07]. Dans le cas d'une expansion combinée à de la sélection naturelle, nous avons vu plus haut que des mutations délétères peuvent elles aussi former des secteurs dans le front d'expansion [HN10]. C'est bien évidemment également le cas pour des mutations avantagées sélectivement, mais les secteurs correspondants sont caractérisés par un angle d'ouverture différent de celui des secteurs associés à des mutations neutres ou délétères [HN10].

Afin de développer des outils permettant de distinguer l'effet de l'expansion de celui de la sélection naturelle, [Hal+07] s'intéresse aux propriétés des frontières entre deux secteurs. L'article montre que ces frontières correspondent à des marches aléatoires superdiffusives, contrairement au cas d'une mutation avantagée sélectivement (en l'absence d'expansion). Cette caractéristique des frontières pourrait donc être utilisée pour faire la distinction entre expansion et sélection. Dans le cas d'une expansion combinée à de la sélection, [HN10] montre via un modèle simple de marches aléatoires modélisant les frontières entre les secteurs et fusionnant en cas de collision qu'il pourrait être possible d'utiliser l'angle d'ouverture des secteurs pour distinguer une mutation neutre d'une mutation (des)avantagée sélectivement.

Si le phénomène d'émergence des secteurs est bien compris qualitativement [ELC04; HN08; KCE06; EFP09] et via l'utilisation de simulations, ce n'est pas le cas d'un point de vue quantitatif. L'approche utilisée dans [Hal+07; HN10] pour étudier les propriétés des secteurs consiste à modéliser leurs frontières par des marches aléatoires qui fusionnent lorsqu'elles se rencontrent, ce qui ne permet d'étudier que le régime post-formation des secteurs. D'un point de vue biologique, la fusion de deux marches aléatoires correspondant aux deux frontières d'un même secteur représente la disparition du secteur. Les principaux résultats théoriques ne présupposant pas l'existence des secteurs [DF16; FE20] portent sur le cas de la dimension 1 uniquement. En dimensions 2 et 3,

[Dur18] considère un modèle similaire à celui étudiée dans [HN08], et étudie de façon relativement informelle la formation de secteurs dans des tumeurs cancéreuses via l'étude des généalogies. Cependant, même si l'étude montre effectivement l'existence de secteurs, ceux-ci sont trop petits pour être observés expérimentalement par biopsie.

D'un point de vue théorique, l'émergence de secteurs dans le front de populations en expansion est donc très mal comprise, ce qui empêche d'utiliser ces secteurs pour faire de l'inférence. L'une des raisons principales à ce manque de résultats théoriques est simple : le manque de modèles stochastiques de génétique des populations adaptés à l'étude des populations en expansion (mais voir par exemple [CM07; DF16; JMW12]). En effet, les modèles classiques de populations en expansion présentent généralement au moins l'un des trois défauts suivants :

- ils sont déterministes, alors que nous avons vu plus haut que l'émergence des secteurs était un phénomène fondamentalement stochastique [HN08];
- ils ne sont définis qu'en dimension 1;
- ils ne sont pas associés à des outils permettant d'étudier la diversité génétique.

Les modèles de génétique des populations, quant à eux, permettent bien de s'intéresser à l'évolution de la diversité génétique, via l'utilisation de processus duaux encodant les généalogies. Cependant, pour construire ces processus duaux, il est souvent nécessaire de supposer que les densités de population sont constantes, ce qui n'est pas le cas dans une population en expansion. Je vais maintenant présenter rapidement plusieurs modèles classiques de populations en expansion. Puis, j'expliquerai quels sont les obstacles à l'utilisation de modèles de génétique des populations pour l'étude de populations en expansion. Je présenterai enfin quels sont les apports de ma thèse pour répondre à cette problématique, et dans quelle mesure les modèles que j'ai construits sont connectés aux modèles classiques de populations en expansion.

Obstacles théoriques à l'étude des secteurs

Les deux grands modèles stochastiques classiques de populations en expansion, que je vais maintenant présenter, sont l'équation de Fisher-KPP [Fis37; KPP37] et le modèle d'Eden [Ede61]. L'équation de Fisher-KPP est à l'origine une équation déterministe modélisant la propagation d'un allèle avantagé sélectivement dans une population uniformément répartie dans tout l'espace. Elle s'est depuis révélée être également applicable à des populations en expansion. La version déterministe a été beaucoup étudiée, y compris en dimension ≥ 2, et il en existe de nombreuses variantes permettant par exemple d'étudier le cas d'individus se déplaçant à des vitesses différentes [Bou+12; Cal+18] ou n'ayant pas les mêmes taux de reproduction [Def+19]. En dimension 1, il est possible d'écrire une version stochastique de l'équation de Fisher-KPP, en ajoutant un terme de bruit de la façon suivante. Si pour tout $t \geq 0$ et $x \in \mathbb{R}$, $p(t,x)$ représente la densité en individus au site x au temps t, alors p est solution de l'équation de Fisher-KPP si pour tout $t \geq 0$ et $x \in \mathbb{R}$,

$$\frac{\partial p}{\partial t}(t,x) = \frac{m}{2}\Delta p(t,x)dt + s_0 p(t,x)(1-p(t,x)) + \sqrt{\frac{1}{p_e}p(t,x)(1-p(t,x))}\dot{W}(dt,dx),$$

où

- m représente un taux de migration,
- s_0 représente le taux de reproduction en l'absence de compétition (c'est à dire, lorsque les densités de population sont très faibles),
- $\dot{W}$ est un bruit blanc espace-temps.

Le terme de bruit correspond à une diffusion de Wright-Fisher (voir la Section 1.2.1), et permet de modéliser la dérive génétique due à la stochasticité dans la reproduction. L'équation de Fisher-KPP admet des solutions en onde progressive [MS95], ce qui permet de comprendre la dynamique de l'expansion au temps long. L'équation n'est cependant pas généralisable au cas de la dimension ≥ 2, car elle n'admet alors plus de solutions du fait du terme de bruit.

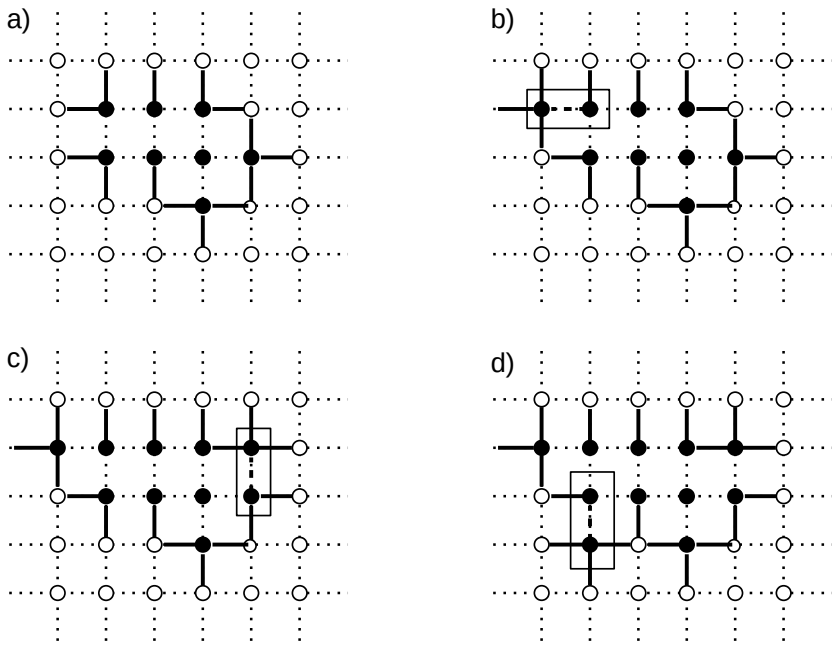


Figure 1.3: Illustration de la dynamique du modèle d'Eden. Les ronds noirs représentent les sommets occupés, et les ronds blancs les sommets vides. Les arêtes en gras sont celles reliant un sommet occupé à un sommet vide. A chaque pas de temps, une arête est choisie uniformément au hasard parmi les arêtes en gras, et le sommet vide correspondant devient occupé.

Le modèle d'Eden, quant à lui, est de nature très différente. Il s'agit d'un modèle discret en temps et en espace, défini sur $\mathbb{Z}^d$, $d \geq 1$, muni de la grille des plus proches voisins. Chaque sommet de la grille est ou bien occupé par une cellule, ou bien vide. Le modèle a initialement été défini de la façon suivante :

- A chaque pas de temps, considérons toutes les arêtes de la grille qui relient un sommet occupé à un sommet vide.
- Choisissons alors l'une de ces arêtes uniformément au hasard.
- Le sommet vide connecté à cette arête devient alors occupé.

Dans le modèle d'Eden, les cellules ne meurent jamais, et sont de plus immobiles. Ainsi, à chaque pas de temps, le nombre de cellules augmente de un, et nous avons bien un modèle d'expansion. Voir la Figure 1.3 pour une illustration de la dynamique.

L'article [JB85] présente deux autres versions de ce modèle : une dans laquelle un sommet vide est choisi uniformément au hasard parmi tous les sommets vides reliés à un sommet occupé, et une dans laquelle un sommet occupé est choisi uniformément au hasard parmi tous les sommets occupés reliés à au moins un sommet vide. Un des (au plus quatre) sommets vides auxquels il est connecté est alors choisi (là encore uniformément au hasard), et devient occupé. Voir la Figure 1.4

pour une illustration de la différence entre le modèle d'Eden originel et les deux variantes que je viens de présenter.

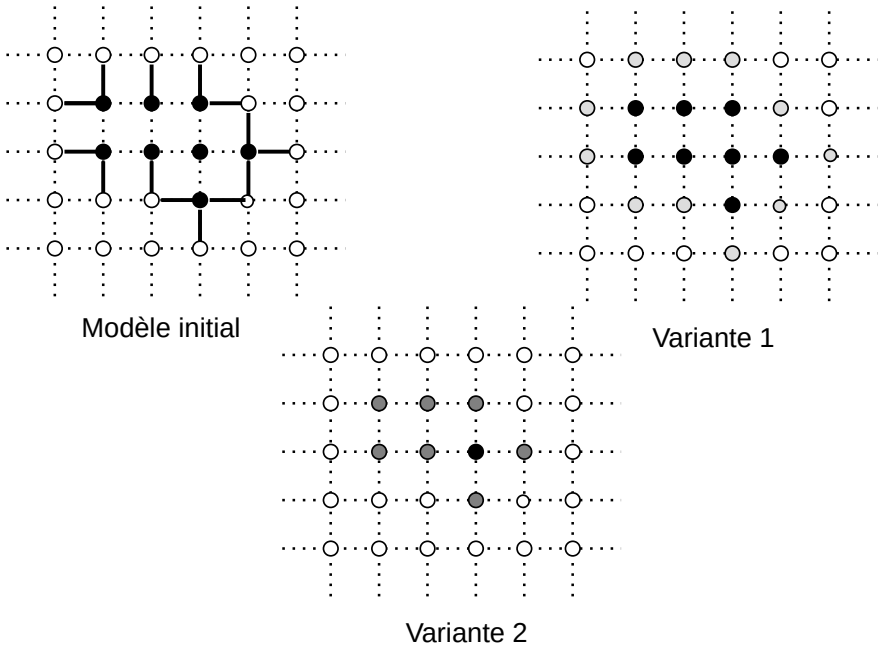


Figure 1.4: Comparaison du modèle d'Eden originel à deux de ses variantes. [Modèle initial] A chaque pas de temps, une arête est choisie uniformément au hasard parmi celles reliant un sommet occupé à un sommet vide (voir arêtes en gras). [Variante 1] A chaque pas de temps, un sommet vide est choisi uniformément au hasard parmi les sommets vides voisins d'un sommet occupé (voir sommets gris clairs). [Variante 2] A chaque pas de temps, un sommet occupé est choisi uniformément au hasard parmi ceux voisins d'un sommet vide (voir sommets gris foncés).

Dans sa version initiale, le modèle d'Eden est ainsi discret en espace et en temps. Il est cependant possible d'en définir une version continue en temps en assignant à chaque arête e un temps de passage $\tau_e \sim \mathcal{E}xp(1)$, tous les temps de passage étant i.i.d. La dynamique devient alors la suivante : si l'un des deux sommets de l'arête devient occupé au temps t, alors l'autre le devient à l'instant $t + \tau_e$, s'il ne devient pas occupé entre temps via l'une des trois autres arêtes auxquelles il est connecté. Ainsi formulé, le modèle d'Eden devient un problème de *percolation de premier passage* [ADH17]. Dans le cas du modèle d'Eden, les temps de passage suivent une loi exponentielle, mais il existe des problèmes de percolation de premier passage beaucoup plus généraux, définis sur des grilles différentes (permettant par exemple des interactions à longue portée) et/ou dans lesquels les temps de passage sont distribués différemment. Tous ces modèles correspondent d'ailleurs à de possibles modèles de populations en expansion, voir par exemple la Section 1.3 de [CD16] et les références qu'elle contient.

Concernant les variantes continues en espace du modèle d'Eden, considérons par exemple celle introduite dans [WLB95]. Dans ce modèle, plus de grille : les cellules sont maintenant représentées par des cercles de rayon $\mathcal{R} > 0$ fixé, qui ne peuvent pas s'intersecter. Il y a donc une notion d'encombrement de l'espace : si une cellule de centre $x \in \mathbb{R}^2$ est présente, alors il ne peut pas y avoir de cellule de centre $y \in \mathbb{R}^2$ satisfaisant $||y - x|| < 2\mathcal{R}$. En particulier, les cellules étant immobiles, une cellule de centre $x \in \mathbb{R}^2$ ne peut se reproduire que s'il existe $y \in \mathbb{R}^2$ tel que $||y-x|| = 2\mathcal{R}$ et tel que pour tout $z \in \mathbb{R}^2 \backslash \{x\}$ correspondant à un centre de cellule, $||z - y|| \geq 2\mathcal{R}$.

La dynamique du processus est alors définie de la façon suivante. A chaque génération, une cellule est choisie uniformément au hasard parmi toutes celles pouvant se reproduire. Un point $y \in \mathbb{R}^2$ satisfaisant la condition introduite plus haut est alors choisi uniformément au hasard parmi tous ceux qui la vérifient, et une cellule de centre y (et de rayon $\mathcal{R}$) est ajoutée au système. Voir la Figure 1.5 pour une illustration de la dynamique.

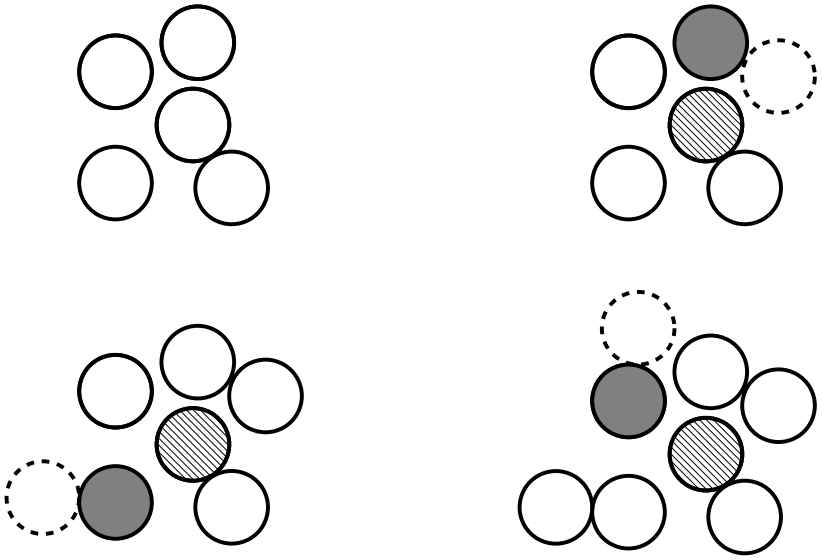


Figure 1.5: Illustration de la dynamique du modèle d'Eden continu en espace. A chaque pas de temps, une cellule (en gris foncé) est choisie uniformément au hasard parmi celles pouvant se reproduire (cellules non hachurées), et donne naissance à une nouvelle cellule.

Si le modèle d'Eden a été autant étudié, c'est en partie parce qu'il est conjecturé comme appartenant à la classe d'universalité de l'équation KPZ (pour Kardar-Parisi-Zhang, voir [KPZ86]). Cette équation qui décrit la croissance de surfaces génère des fronts d'expansion qui semblent avoir les mêmes propriétés que ceux observés dans des populations réelles en expansion (voir par exemple [Hue+10]). D'autres modèles de croissance sont eux aussi conjecturés comme appartenant à cette classe d'universalité, mais il s'agit d'une conjecture difficile à prouver, qui n'a pu être démontrée rigoureusement que pour le modèle de croissance Solid-On-Solid [BG97].

Ainsi, le modèle d'Eden est considéré comme un bon candidat pour modéliser les populations en expansion. Cependant, et comme pour beaucoup de modèles issus de la théorie de la percolation, il est difficile d'obtenir des résultats théoriques autrement qu'en utilisant des simulations. Il semble ainsi compliqué d'étudier l'émergence de secteurs dans le modèle d'Eden, dans le cas des variantes discrètes comme continues en espace, du fait du manque d'outils associés permettant d'étudier la diversité génétique. L'équation de Fisher-KPP ne peut pas être utilisée non plus, car sa version stochastique n'admet pas de solutions en dimension ≥ 2.

Du côté de la génétique des populations, il existe de nombreux modèles stochastiques pour lesquels il est possible d'étudier la diversité génétique. Dans la section suivante (Section 1.2), j'en présente deux d'entre eux : le modèle de Wright-Fisher et le processus Λ-Fleming Viot spatial (ou SLFV). Voir aussi [Eth11] pour une introduction plus complète au modèle de Wright-Fisher et à d'autres modèles de génétique des populations.

En génétique des populations, une façon classique d'étudier la diversité génétique consiste à remonter dans le temps et à reconstituer les généalogies d'un échantillon d'individus. En effet, dans des populations haploïdes et en l'absence de mutations, deux individus ayant un ancêtre commun ont le même type. De même, si des mutations sont possibles, les seules différences qui peuvent exister entre deux individus ayant un ancêtre commun sont liées à des mutations s'étant produites depuis cet ancêtre. De ce fait, il est possible de distinguer deux grands types de questions liées à la diversité génétique :

- Celles liées à l'évolution de la diversité génétique en l'absence de mutations, et sachant la structure initiale de la diversité.

C'est dans cette catégorie que rentre l'étude de l'émergence de secteurs dans le front de populations en expansion.

- Celles liées à la propagation de mutations dans la population, et au nombre de mutations différentes que l'on s'attend à trouver dans un échantillon d'individus.

Ceci est motivé par des problématiques d'inférence à partir de données génétiques réelles. C'est principalement dans cette catégorie que rentrerait l'étude des angles d'ouverture des secteurs correspondant à des mutations (des)avantagées sélectivement. Cependant, il est souvent supposé que les mutations considérées sont neutres.

Même si la plupart des modèles de génétique des populations classiques sont sans structuration spatiale, il est généralement possible d'introduire une composante spatiale (à condition d'éviter le problème de "pain in the torus" [Fel75], c'est à dire d'explosions locales du nombre d'individus en l'absence de contrôle des densités locales), soit en divisant l'espace en petites zones appelées *dèmes* ou *patchs*, dans lesquelles la structuration spatiale peut être négligée (voir la Section 1.2.2), soit à la façon du processus Λ-Fleming Viot spatial, en tirant aléatoirement des zones de l'espace dans lesquelles la reproduction a lieu (voir la Section 1.2.3). La dimension 2 ou plus et la nécessité d'outils permettant d'étudier la diversité génétique ne sont donc a priori pas un problème en génétique des populations. En revanche, il est généralement nécessaire de supposer des tailles de population constantes. En effet, grâce à cette hypothèse, la compétition entre individus est d'intensité constante, ce qui permet de remonter dans le temps et de reconstituer les généalogies d'un échantillon sans avoir à chercher avec combien d'individus ses ancêtres étaient en interaction. Cette hypothèse n'est cependant pas du tout adaptée au cas de populations en expansion.

Ainsi, le but de ma thèse est de développer des modèles de génétique des populations structurés en espace, adaptés aux populations en expansion et permettant d'étudier les patterns de diversité génétique au niveau du front d'expansion. Ces modèles sont un prérequis indispensable à l'étude théorique de l'émergence et des propriétés des secteurs observés expérimentalement. L'approche que j'ai utilisée consiste à intégrer à différents modèles de génétique des populations la technique introduite dans [DF16] et [HN08]. Basée sur la théorie des systèmes de particules en interaction, elle consiste à remplir les zones vides avec des individus "fantômes", qui peuvent se reproduire, mais avec un fort désavantage sélectif face aux individus réels. La reproduction des individus fantômes permet de modéliser les fluctuations stochastiques des densités de population, et de ne pas avoir à définir à l'avance la dynamique de l'expansion. Voir la Section 1.3 pour une présentation plus en détail de cette approche, et de comment je l'ai appliquée au processus Λ-Fleming Viot spatial et à une variante du modèle de Wright-Fisher.

1.1.2 Propagation d'espèces végétales en milieu urbain via les pieds d'arbres

Pourquoi s'intéresser aux pieds d'arbres d'alignement ?

La deuxième question à laquelle je me suis intéressée concerne la dynamique des plantes herbacées sauvages poussant dans un environnement très particulier : les pieds d'arbres d'alignement, qui se

trouvent le long des rues en ville (voir la Figure 1.6 pour une illustration). Ces pieds d'arbres sont très petits, de l'ordre du mètre carré, et sont généralement entourés de goudron. Ainsi, les plantes ne peuvent généralement pas pousser en dehors des pieds d'arbres. En plus d'être un environnement très fragmenté, les pieds d'arbres constituent aussi un environnement perturbé, dans lequel les événements d'extinction sont fréquents. Ils sont ainsi régulièrement piétinés, désherbés,...



Figure 1.6: (a) Pied d'arbre d'alignement près de Bordeaux, France. (b) Pied d'arbre d'alignement dans Bayonne, France.
(c) Pieds d'arbres d'alignement près de Bayonne, France.

L'intérêt de l'étude de ce système est double : écologique comme théorique. D'un point de vue écologique tout d'abord, les pieds d'arbres d'alignement peuvent servir de corridors écologiques entre des espaces verts plus grands, tels que des parcs et des jardins. Ils peuvent ainsi potentiellement s'inscrire dans la trame verte en milieu urbain et contribuer à la connectivité globale, qui améliore la qualité de l'écosystème urbain. Ce n'est cependant pas forcément le cas en pratique, du fait des perturbations fréquentes conduisant à des extinctions plus ou moins localisées. Il est donc pertinent de chercher à comprendre dans quelle mesure les pieds d'arbres servent effectivement de corridors écologiques, et si oui, pour quelles espèces. Pour les espèces pour lesquelles ce n'est pas le cas, une étude théorique peut permettre de comprendre comment modifier les méthodes de gestion des pieds d'arbres afin de rendre ceci possible.

De plus, d'un point de vue théorique, les pieds d'arbres d'une rue sont très adaptés à la modélisation mathématique. En effet, ils sont bien délimités, d'une taille similaire, disposés le long d'une ligne, et souvent équidistants. Par ailleurs, du fait du passage régulier des jardiniers pour désherber les pieds d'arbres, une plante ne peut généralement pas vivre plus d'un an, de façon similaire aux espèces annuelles. Les générations peuvent ainsi être considérées comme non chevauchantes. Toutes ces caractéristiques correspondent à des hypothèses souvent nécessaires pour pouvoir faire une analyse mathématique, et qui sont ici effectivement vérifiées.

Plusieurs études ont montré que les pieds d'arbres de certaines rues des villes de Paris [Oma+19] ou Montpellier [DPC11] correspondaient effectivement à des corridors écologiques pour certaines espèces de plantes. Celles-ci s'échappent des différents espaces verts (parcs, jardins,...) sous forme de graines, et colonisent plus ou moins rapidement les rues avoisinantes génération après génération. Ceci pourrait être à double tranchant : en effet, du fait des perturbations fréquentes qu'ils subissent, les pieds d'arbres pourraient faciliter la propagation d'espèces invasives [ABH00], comme le font par exemple les voies ferrées pour le sénéçon du Cap (*Senecio inaequidens*) dans Paris [Bla+15a]. De façon plus générale, l'environnement urbain, très perturbé, est plutôt favorable à certaines espèces exotiques (i.e, aux espèces introduites délibérément ou non par l'homme) [Mur+07; Pyš98], dont les espèces invasives sont un sous-ensemble particulièrement dynamique. Ceci est sans doute au moins en partie liée à la présence de parcs et jardins, connus pour abriter un grand nombre d'espèces exotiques et former des réservoirs potentiels [Deh+07a; Deh+07b; Smi+06], et est d'autant plus vrai que l'urbanisation est forte [Kow95; MPS00; MP90; Mur+07]. Ainsi, les travaux de thèse de Noélie Maurel [Mau10], portant sur la distribution des espèces végétales en Île de France, ont mis en évidence un plus grand nombre d'espèces invasives dans les

milieux urbains (principalement les villes, les petits parcs et les anciens sites industriels) et agricoles que dans les milieux forestiers ou ouverts.

Pourtant, si les pieds d'arbres abritent eux aussi un grand nombre d'espèces exotiques [Mau10], ils contiennent en fait relativement peu d'espèces invasives, ce qui est à première vue surprenant étant donné le caractère très perturbé de ce milieu. L'étude menée dans [Che+08] apporte des éléments de réponse : les espèces invasives sont généralement des espèces caractérisées (entre autres) par une forte dispersion [Sak+01; WG04]. Or, les pieds d'arbres forment un environnement très fragmenté, une forte dispersion signifie ici une probabilité élevée de voir les graines tomber hors des pieds d'arbres, là où elles ne peuvent pas germer. Il peut donc y avoir de la sélection contre la dispersion, comme ceci a été observé pour le crépis de Nîmes (*Crepis sancta*) dans des pieds d'arbres de Montpellier [Che+08]. Les espèces qui sont invasives dans d'autres environnements peuvent ainsi ne pas être adaptées à la survie dans les pieds d'arbres. Ceci illustre d'ailleurs les résultats de [ABH00], qui montre que les caractéristiques qui rendent une espèce invasive sont très écosystème-dépendantes.

Modèles de banque de graines

Si l'étude menée dans [Che+08] permet de comprendre pourquoi les espèces invasives sont peu représentées dans la flore des pieds d'arbres, elle soulève aussi des questions. En effet, si la dispersion est contre-sélectionnée et si les événements d'extinction sont fréquents, comment des espèces arrivent-elles à se maintenir dans les pieds d'arbres ? Une possibilité est l'apport continu de nouvelles graines depuis une source extérieure, par exemple les parcs et jardins. Une autre pourrait être la présence d'une banque de graines dans le sol, c'est à dire d'un stock de graines viables mais dormantes pouvant potentiellement germer plusieurs générations après leur production. L'étude menée dans [Oma+19] suggère en effet une influence de la banque de graines sur la dynamique de certaines espèces poussant dans les pieds d'arbres. Le vérifier en pratique, par exemple en prélevant des échantillons de sol et en les plaçant dans des chambres de germination [BB14], est cependant difficile à mettre en œuvre. En effet, en plus des difficultés générales associées à cette approche, se pose aussi le problème du fort tassement du sol (qui est parfois trop dur pour creuser facilement), ou de la présence de déchets et déjections qui rendent la collecte d'échantillons difficile. Ceci encourage à se tourner vers la modélisation, afin de mener des analyses théoriques et de faire de l'inférence à partir de données réelles.

Les modèles de banque de graines peuvent être divisés en deux grandes catégories : des modèles plutôt orientés génétique des populations, visant à étudier l'effet d'une banque de graines sur la diversité génétique et à faire de l'inférence à partir de données génétiques, et des modèles orientés dynamique des populations, s'intéressant plutôt à l'évolution du nombre d'individus ou de la présence/absence de l'espèce. Dans les deux cas, il s'agit de modèles relativement récents.

Du côté de la génétique des populations, la plupart des modèles se basent sur celui introduit dans [KKL01]. Cet article incorpore une banque de graines dans le modèle de Wright-Fisher, un modèle classique de génétique des populations. Voir les Sections 1.2.1 et 1.2.2 pour une présentation du modèle de Wright-Fisher et de l'extension au cas d'une banque de graines introduite dans [KKL01]. Dans le modèle initial de [KKL01], la durée de vie des graines avant perte de viabilité est bornée, mais cette hypothèse a depuis été relâchée [Bla+13; Bla+16]. Les modèles de [Bla+16] et [KKL01] ont servi de base à d'autres modèles de banque de graines incorporant de la sélection [Koo+17] ou une structuration spatiale [HP17; HN21; GHO22]. Dans certains cas, il est possible de faire de l'inférence à partir de données génétiques [Bla+20; Sel+20; Tel+11], afin de détecter la présence d'une banque de graines et/ou de la prendre en compte dans l'estimation d'autres paramètres.

Si les modèles issus de la génétique des populations supposent souvent des tailles de popula-

tion constantes, ce n'est pas le cas des modèles plus orientés dynamique des populations, qui permettent de suivre l'évolution du nombre d'individus [Jar+95; LCP19] ou de la présence/absence de l'espèce [AP01; Bor+15; FW02; Plu+18] (voir l'introduction de [LCP19] pour une liste de références plus exhaustive). Ayant à disposition des données de présence/absence, je me suis plus spécifiquement intéressée aux modèles de type SPOM (*Stochastic Patch Occupancy Models*, voir la Section 1.4.1), qui sont particulièrement adaptés à l'étude des pieds d'arbres [DPC11; Oma+19] et permettent dans certains cas de faire de l'inférence [Kaz+21; Plu+18].

Apports de ma thèse

Dans ma thèse, je me suis intéressée au rôle joué par une potentielle banque de graines sur la dynamique des plantes dans les pieds d'arbres, et j'ai cherché à comprendre dans quelle mesure la capacité à former une banque de graines pouvait être sélectionnée dans ce type d'environnement. J'ai abordé la question sous deux angles :

- Un angle plus théorique, passant par le développement d'un modèle pour la dynamique des plantes dans les pieds d'arbres qui incorpore une banque de graines.

Ce modèle est basé sur des idées issues de plusieurs modèles différents afin de prendre en compte la banque de graines [Bla+16; KKL01] et les événements d'extinction locaux [DF16; HN08], et est adapté à toute métapopulation de plantes dans un environnement fragmenté caractérisé par des événements d'extinction locaux fréquents. Dans ma thèse, j'utilise ce modèle pour montrer l'existence d'une probabilité d'extinction critique au delà de laquelle une banque de graines est nécessaire pour survivre et se propager (la preuve reposant sur un argument de percolation), ainsi que pour valider théoriquement un modèle simplifié de dynamique des populations de type SPOM plus simple à étudier. Cependant, il pourrait aussi être utilisé pour étudier la diversité génétique dans ces métapopulations (voir la Section 1.3.3). Je présente plus en détail les résultats obtenus dans la Section 1.3.3 et le Chapitre 6.

- Un angle plus pratique, basé sur de l'inférence à partir de données réelles.

Pour cela, j'ai pu m'appuyer sur un jeu de données de 10 ans d'inventaires floristiques réalisés de 2009 à 2018 sur 1324 pieds d'arbres du 12ème arrondissement de Paris par l'équipe de Nathalie Machon [Mac20]. En me basant sur le modèle et l'estimateur introduits dans [Plu+18], j'ai défini une métrique qui permet de mesurer la contribution d'une potentielle banque de graines à la dynamique observée. L'application de cette métrique au jeu de données a permis de confirmer les résultats de [Oma+19]. De plus, l'étude des performances de l'estimateur associé montre que cette métrique peut être appliquée à une plus grande variété de jeux de données, comme par exemple des jeux de données issus des sciences participatives. Voir la Section 1.4.2 pour l'idée générale derrière la métrique, et voir le Chapitre 5 pour la définition et l'étude de cette métrique.

1.2 Quelques modèles de génétique des populations pour des populations à l'équilibre

Les modèles de génétique des populations que j'ai construits et étudiés dans le cadre de ma thèse sont basés sur deux modèles classiques : le modèle de Wright-Fisher, développé indépendamment par R. Fisher [Fis58] et S. Wright [Wri31], et le processus Λ-Fleming Viot spatial [BEV10; Eth08], introduit à la fin des années 2000. Dans sa version d'origine, le modèle de Wright-Fisher est sans structuration spatiale, et suppose que la population étudiée est composée d'un nombre fini et constant d'individus, noté N. Il s'agit de plus d'un modèle à temps discret : à chaque pas de temps,

correspondant à une génération, les N individus qui composent la population se reproduisent, donnent naissance à N descendants, puis meurent.

Le processus Λ-Fleming Viot spatial, quant à lui, est à temps continu, et suppose qu'une infinité d'individus sont uniformément répartis dans un espace $E \subseteq \mathbb{R}^d$, $d \geq 1$, qui peut être pris compact ou non. Comme son nom le suggère, il comporte une structuration spatiale, au sens où les descendants d'un individu sont proches de lui géographiquement.

En dehors de ces différences fondamentales, le modèle de Wright-Fisher et le processus Λ-Fleming Viot spatial possèdent plusieurs caractéristiques communes :

- Ils décrivent des populations d'individus *haploïdes*, c'est-à-dire des populations d'individus possédant une seule copie de chaque chromosome, et qui n'ont donc qu'un seul parent.
- Dans leurs versions originelles, ils permettent d'étudier l'évolution du nombre (modèle de Wright-Fisher) ou de la densité (processus Λ-Fleming Viot spatial) d'individus d'un type donné dans une population comportant plusieurs types neutres sélectivement.
- La façon dont ces modèles sont définis permet de les modifier facilement afin d'intégrer de nouvelles composantes : par exemple, de la sélection naturelle (voir les Sections 1.2.2 et 1.2.3), une banque de graines (voir la Section 1.2.2), des événements d'extinction locaux (voir la Section 1.2.2),...

Le fait que ces deux modèles soient adaptés aux populations haploïdes n'est pas restrictif. En effet, il est généralement possible de se ramener au cas de populations haploïdes en raisonnant chromosome par chromosome plutôt qu'individu par individu.

Dans toute la suite, afin de simplifier les notations, nous considérerons que la population n'est composée que de deux types (neutres sélectivement), notés A et a. Il est toutefois possible de généraliser les définitions à un nombre plus grand de types. Dans le modèle de Wright-Fisher, le nombre d'individus de type a à la génération $n \in \mathbb{N}$ sera noté $X(n) \in [\![0, N]\!]$. Le processus Λ-Fleming Viot spatial est quant à lui un processus à valeurs mesures, mais nous pouvons considérer de façon informelle qu'il décrit la densité en individus de type a en chaque point de l'espace et à chaque instant. Aussi, nous noterons $\omega_t(x) \in [0, 1]$ la proportion d'individus de type a en $x \in E$ au temps $t \geq 0$.

1.2.1 Le modèle de Wright-Fisher

Présentation du modèle

Cette section est adaptée de [Eth11]. S'y référer pour les preuves des différents résultats présentés ici.

Commençons par définir rigoureusement le modèle de Wright-Fisher. Pour cela, supposons qu'initialement, le nombre $X(0)$ d'individus de type a est égal à $X^0 \in [\![0, N]\!]$. La dynamique du modèle peut alors être décrite de la façon suivante. Sachant le nombre $X(n)$ d'individus de type a au début de la génération n, le nombre $X(n+1)$ d'individus de type a au début de la génération suivante suit une loi binomiale

$$X(n+1) \sim \mathrm{Binom}(N, X(n)N^{-1}).$$

Cela correspond à considérer que chacun des N descendants choisit un parent uniformément au hasard, et adopte son type. Voir la Figure 1.7 pour une illustration de la dynamique.

Sous cette définition,

$$\mathbb{E}\left[X(n+1)|X(n)\right] = X(n).$$

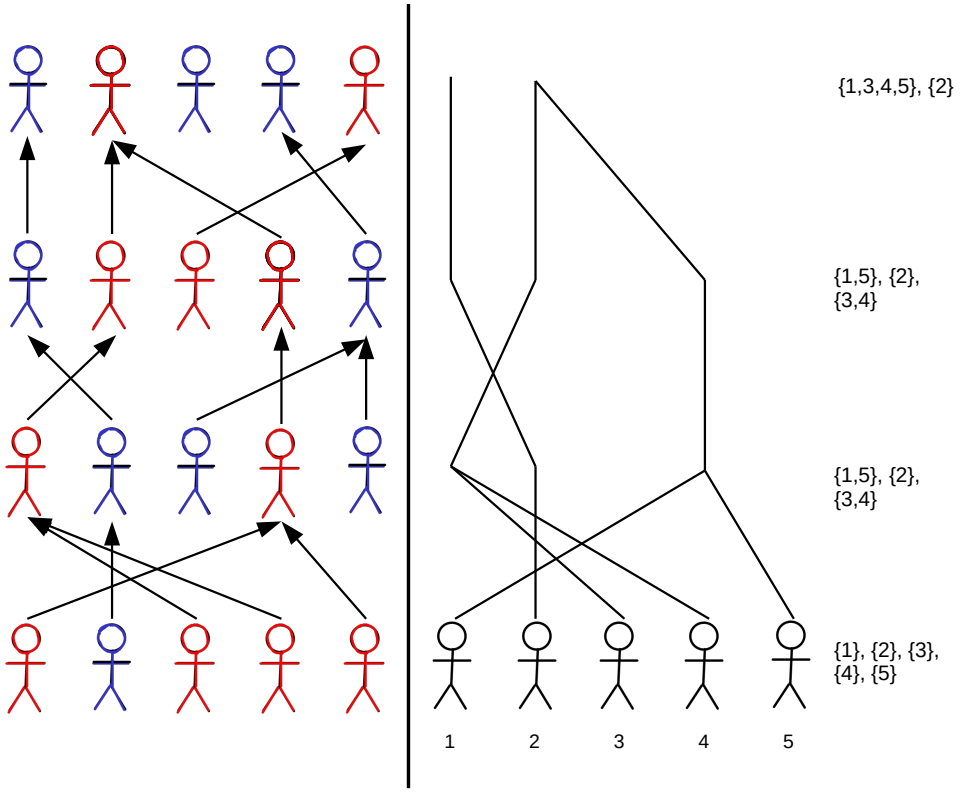


Figure 1.7: Illustration de la dynamique du modèle de Wright-Fisher, et du lien avec le m-coalescent discret encodant les généalogies. Les individus bleus correspondent aux individus de type A, et les individus rouges aux individus de type a.

Ainsi, les types a et A sont bien neutres sélectivement, et le nombre d'individus de type a est en moyenne constant. Cependant,

$$\mathrm{Var}(X(n+1)|X(n)) = X(n)(N - X(n))N^{-1} \neq 0$$

sauf si l'un des deux types est éteint. Le nombre d'individus de type a peut donc fluctuer du fait de la stochasticité dans la reproduction, un phénomène appelé *dérive génétique*. A terme, ces fluctuations conduiront à l'extinction de l'un des deux types et à la fixation de l'autre.

Remarque 1.2.1. Il existe une variante en temps continu du modèle de Wright-Fisher, appelée *modèle de Moran* [Mor58]. Dans ce modèle, au lieu d'une reproduction par génération, le temps entre deux événements de reproduction successifs est cette fois-ci donné par une loi exponentielle de paramètre $\binom{N}{2}$, où N est le nombre d'individus total. A chaque événement de reproduction, une paire d'individus est choisie uniformément au hasard dans la population. L'un des deux individus meurt, tandis que l'autre se reproduit, et son descendant remplace l'individu qui vient de mourir.

Processus dual

Pour comprendre l'évolution de la diversité génétique et faire de l'inférence à partir de données réelles, il est nécessaire de s'intéresser à la distribution du nombre de types différents dans un échantillon de $m \in [\![2, N]\!]$ individus au temps $n \in \mathbb{N}$. Afin de répondre à cette question, il est possible d'envisager deux approches :

- Une approche *forwards-in-time*, qui consiste à travailler avec la distribution explicite de $X(n)$, ou à simuler l'ensemble de la population jusqu'au temps n.

L'inconvénient de cette approche est double. Tout d'abord, la distribution de $X(n)$ devient très vite difficile à calculer. De plus, d'un point de vue computationnel, ne sachant pas à l'avance quelle fraction de la population contient les ancêtres de l'échantillon d'intérêt, il est nécessaire de simuler la dynamique de l'ensemble des N individus jusqu'à la génération n. Ceci peut être difficile pour de très grandes valeurs de N, proches de la limite en grande population.

- Une approche *backwards-in-time*, basée sur un processus dual encodant les généalogies des individus de l'échantillon.

L'idée derrière cette approche est la suivante : en l'absence de mutations et dans une population d'individus haploïdes, le type d'un individu pris à la génération n est le même que celui de son ancêtre au temps $n-1$, $n-2$, ..., 0. De plus, si deux individus ont un ancêtre commun, alors ils ont le même type. La généalogie des individus de l'échantillon encode donc les corrélations entre les types de ces individus.

Nous allons ainsi nous intéresser au processus $(\Pi_n)_{n\in\mathbb{N}}$ qui décrit les généalogies dans un échantillon de m individus, processus que nous appellerons le *m-coalescent discret*. Ce processus est défini sur l'ensemble $\mathcal{P}_m$ des partitions de $[\![1,m]\!]$, de la façon suivante.

Définition 1.2.2. *(m-coalescent discret) Le m-coalescent discret $(\Pi_n)_{n\in\mathbb{N}}$ est la chaîne de Markov à temps discret définie sur l'espace d'états $\mathcal{P}_m$ et de condition initiale $\{\{1\},\{2\},...\{m\}\}$ dont les transitions sont les suivantes : pour tout $n\in\mathbb{N}$, si Π_n est composée de $|\Pi_n|>1$ blocs, alors chaque bloc tire (indépendamment des autres) un entier uniformément au hasard dans $[\![1,N]\!]$, et les blocs ayant tiré le même entier fusionnent. Si Π_n est composée d'un seul bloc, alors $\Pi_{n+1}=\Pi_n$.*

Ce processus peut être interprété comme encodant les généalogies dans le modèle de Wright-Fisher. En effet, numérotons par 1, 2, ... les m individus d'un échantillon dans une population de N individus évoluant selon le modèle de Wright-Fisher, et intéressons-nous aux parents de ces individus. Chacun d'entre eux est choisi uniformément au hasard parmi les N individus de la génération précédente, ce qui correspond à tirer un entier uniformément au hasard dans $[\![1,N]\!]$. Une fusion de blocs dans le m-coalescent correspond alors à plusieurs individus qui trouvent un ancêtre commun. Ainsi, pour tout $0\leq n'\leq n$, chaque bloc de $\Pi_{n'}$ peut être interprété commun un ensemble d'individus ayant le même ancêtre à la génération $n-n'$, et $|\Pi_{n'}|$ correspond au nombre d'ancêtres différents de l'échantillon à la génération $n-n'$. En particulier, si $|\Pi_{n'}|=1$, alors tous les individus descendent du même ancêtre dans la génération $n-n'$.

Il est de plus possible d'utiliser ce processus pour en déduire la distribution des types dans l'échantillon au temps présent. En effet, le cas $n'=n$ correspond à la génération initiale, et il ne reste plus qu'à regarder le type de chaque ancêtre dans la condition initiale pour conclure. Par exemple, si $|\Pi_n|=1$, alors nous pourrons en déduire que les m individus de l'échantillon au temps n ont tous le même type, et sont de type a avec probabilité X^0N^{-1}. Voir la Figure 1.7 pour une illustration.

Formellement, le modèle de Wright-Fisher et le m-coalescent discret satisfont la relation de dualité suivante.

Proposition 1.2.3. *Soient $m\in[\![2,N]\!]$ et $X^0\in[\![0,N]\!]$. Si $(X(n))_{n\geq 0}$ est le modèle de Wright-Fisher de condition initiale X^0 et si $(\Pi_n)_{n\geq 0}$ est le m-coalescent discret, alors pour tout $n\in\mathbb{N}$,*

$$\mathbb{E}_{X^0}\left[\frac{\binom{X(n)}{m}}{\binom{N}{m}}\right]=\mathbb{E}_{\{\{1\},\{2\},...,\{m\}\}}\left[\frac{\binom{X^0}{|\Pi_n|}}{\binom{N}{|\Pi_n|}}\right],$$

ce qui peut aussi s'écrire

$$\mathbb{E}_{X^0}\left[\frac{\binom{X(n)}{|\Pi_0|}}{\binom{N}{|\Pi_0|}}\right]=\mathbb{E}_{\{\{1\},\{2\},...,\{m\}\}}\left[\frac{\binom{X(0)}{|\Pi_n|}}{\binom{N}{|\Pi_n|}}\right].$$

En d'autres termes, les m individus d'un échantillon choisis uniformément au hasard parmi les N individus de la population au temps n sont tous de type a si, et seulement si tous leurs ancêtres au temps 0 sont de type a.

D'un point de vue computationnel, l'intérêt du processus dual vient du fait qu'il permet de restreindre la simulation du processus à un échantillon de taille au plus m, ce qui permet de s'approcher facilement de la limite en grande population $N \to +\infty$.

Limite en grande population

Nous avons vu précédemment que pour tout $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[X(n+1) - X(n) | X(n)\right] &= 0 \\ \text{et } \mathrm{Var}\left[X(n+1)|X(n)\right] &= X(n)(N - X(n))N^{-1},\end{aligned}$$

ce qui implique que les fluctuations du nombre d'individus de type a sont uniquement dues à la stochasticité dans la reproduction. De plus, nous pouvons déduire du processus dual que le temps moyen pour que deux individus se trouvent un ancêtre commun est égal à N.

Lorsque N devient grand, la variance de la *proportion* d'individus de type a vérifie

$$\mathrm{Var}\left[\frac{X(n+1)}{N}\middle|\frac{X(n)}{N}\right] = \frac{X(n)}{N}\left(1 - \frac{X(n)}{N}\right)N^{-1} \propto N^{-1},$$

et il est donc nécessaire de changer d'échelle de temps et compter en unité de N générations pour observer les fluctuations dans la limite $N \to +\infty$. C'est aussi le changement d'échelle à faire pour pouvoir observer la fixation de l'un des deux types : sur une échelle de temps plus courte, les individus ne se trouveraient jamais d'ancêtre commun, et sur une échelle de temps plus longue, la fixation d'un type aurait lieu instantanément.

Le processus limite obtenu en prenant $N \to +\infty$ et en faisant le changement d'échelle de temps décrit plus haut est appelé *diffusion de Wright-Fisher*, et est défini de la façon suivante.

Définition 1.2.4. *(Diffusion de Wright-Fisher) Soit $p^0 \in [0,1]$. La diffusion de Wright-Fisher $(p_t)_{t\geq 0}$ de condition initiale p^0 est l'unique solution de l'équation différentielle stochastique*

$$\begin{cases} dp_t &= \sqrt{p_t(1-p_t)}dB_t \\ p_0 &= p^0, \end{cases}$$

où $(B_t)_{t\geq 0}$ est un mouvement brownien.

La diffusion de Wright-Fisher est elle aussi associée à un processus dual encodant les généalogies, qui est la limite du m-coalescent lorsque $N \to +\infty$. En effet, au temps n, si le m-coalescent Π_n contient au moins deux blocs, alors entre les temps n et $n+1$:

- une seule paire de blocs fusionne avec probabilité $\propto N^{-1}$,
- une seule fusion d'au moins trois blocs se produit avec probabilité $\propto N^{-2}$,
- au moins deux fusions d'au moins deux blocs se produisent avec probabilité $\propto N^{-2}$.

En d'autres termes, si nous changeons d'échelle de temps et comptons en unités de N générations, les seules transitions visibles sont les fusions d'exactement une paire de blocs. Les fusions simultanées ne sont pas visibles dans cette échelle de temps. Le processus obtenu est appelé *coalescent de Kingman*, et est défini de la façon suivante.

Définition 1.2.5. *(Coalescent de Kingman) Soit $m \in \mathbb{N}\backslash\{0,1\}$. Le coalescent de Kingman de taille d'échantillon m est la chaîne de Markov à temps continu $(\pi_t)_{t\geq 0}$ définie sur l'espace d'états $\mathcal{P}_m$ dans laquelle chaque paire de blocs fusionne indépendamment des autres à taux 1.*

La diffusion de Wright-Fisher et le coalescent de Kingman satisfont la relation de dualité suivante.

Proposition 1.2.6. *Soient $p^0 \in [0,1]$ et $m \in \mathbb{N}\backslash\{0,1\}$. Soient $(p_t)_{t\geq 0}$ la diffusion de Wright-Fisher de condition initiale p^0 et $(\pi_t)_{t\geq 0}$ le coalescent de Kingman de taille d'échantillon m. Alors, $(p_t)_{t\geq 0}$ et $(\pi_t)_{t\geq 0}$ satisfont la relation de dualité*

$$\forall t \geq 0, \quad \mathbb{E}_{p^0}\left[p_t^{|\pi_0|}\right] = \mathbb{E}_{\{\{1\},\{2\},\ldots,\{m\}\}}\left[p_0^{|\pi_t|}\right],$$

qui peut aussi s'écrire

$$\forall t \geq 0, \quad \mathbb{E}_{p^0}\left[p_t^m\right] = \mathbb{E}_{\{\{1\},\{2\},\ldots,\{m\}\}}\left[(p^0)^{|\pi_t|}\right].$$

1.2.2 Variations sur le modèle de Wright-Fisher

Ajout d'une structuration spatiale

Dans sa version originelle, le modèle de Wright-Fisher ne comporte pas de structuration spatiale. Il en existe cependant de nombreuses variantes, qui rajoutent une composante spatiale au modèle tout en conservant l'existence d'une relation de dualité. L'approche utilisée consiste à diviser l'espace en sous-unités, appelées *dèmes* ou *patchs*, qui sont suffisamment petites pour pouvoir négliger la structuration spatiale à l'intérieur. Dans chaque patch, la reproduction suit un modèle de Wright-Fisher, et des migrations sont possibles d'un patch à l'autre. Afin de garder une taille de patch constante, hypothèse nécessaire pour pouvoir définir le processus dual, les migrations peuvent être modélisées de deux façons :

- En imposant un nombre d'individus échangés entre chaque paire de patchs et à chaque génération. Les individus migrants sont alors choisis uniformément au hasard dans chaque patch, indépendamment de leur type.
- En intégrant les migrations à la phase de choix des parents dans le modèle de Wright-Fisher. Un individu dans le patch i a alors une probabilité $m_{i,j}$ de choisir son parent dans le patch j.

Il est possible de distinguer deux grandes familles de modèles structurés spatialement : les *modèles d'îles* et les *modèles stepping-stone*, ou *modèles en pas japonais*. La différence entre ces deux familles de modèles provient de la forme de la structuration spatiale : les migrations d'un patch i à un patch j sont toujours possibles dans un modèle d'îles, mais pas forcément dans un modèle stepping-stone. Elles dépendent alors par exemple de la distance entre les deux patchs.

Remarque 1.2.7. A noter que ces deux familles de modèles ne contiennent pas que des modèles de Wright-Fisher structurés spatialement, et incluent aussi d'autres modèles de génétique des populations basés sur des modèles de Moran, des diffusions de Wright-Fisher,...

Intéressons-nous d'abord aux modèles d'îles. Basés sur le modèle d'îles de Wright introduit dans [Wri31], ils consistent à diviser la population en $I \in \mathbb{N} \cup \{\infty\}$ communautés, aussi appelées *îles* (d'où le nom du modèle), l'île d'indice $i \in [\![1, I]\!]$ contenant $N_i = \mathbb{N}\backslash\{0\}$ individus. D'un point de vue historique, [Wri31] considère un nombre infini de patchs, tandis que [Lat73; Mar70] se sont intéressées les premiers à un nombre fini d'îles.

Considérons qu'à chaque génération et dans chaque patch d'indice $i \in [\![1, I]\!]$, un nombre $M_i \in [\![1, N_i]\!]$ d'individus migrent hors du patch. Les migrants qui proviennent de chacun des patchs forment alors un *réservoir* d'individus. Puis, chaque patch est complété en choisissant M_i individus dans le réservoir, indépendamment de leur origine. Ceci correspond à une structure spatiale assez faible, au sens où la distance entre les patchs n'intervient nulle part. Pour autant, tous les patchs ne contribuent pas forcément au réservoir d'individus de façon égale. Par exemple, un individu choisi au hasard dans le réservoir a une probabilité $M_i \left(\sum_{i'=1}^{I} M_{i'}\right)^{-1}$ de provenir du patch i. Ainsi, les probabilités de migration d'un individu du patch i vers les patchs $j \neq j'$ ne sont généralement pas égales.

Les modèles d'îles sont donc basés sur cette idée de réservoir d'individus auquel chaque patch contribue, réservoir qui est ensuite redistribué entre les patchs indépendamment de l'origine des individus. Si les modèles de type *stepping-stone* considèrent eux aussi que la population est divisée en $I \in \mathbb{N} \cup \{\infty\}$ patchs, ils supposent de plus que ces patchs correspondent aux sommets d'un graphe $\mathcal{G} = (V, E)$, $|V| = I$ qui encode les migrations possibles. Le cas généralement considéré est celui où $\mathcal{G}$ est le graphe des plus proches voisins sur $\mathbb{Z}^2$ ou $\mathbb{Z}$.

Remarque 1.2.8. A noter que les modèles d'îles correspondent en fait au cas particulier du modèle stepping-stone associé à un graphe $\mathcal{G}$ complet.

Le premier modèle de type stepping-stone a été introduit par Kimura dans [Kim53]. Ce modèle se place dans la limite en grande population du modèle de Wright-Fisher, et décrit l'évolution des fréquences $(p_i)_{i\in V}$ d'individus de type a dans chaque patch, à partir de diffusions de Wright-Fisher. Ainsi, pour tout $i \in V$, p_i est solution de l'équation différentielle stochastique

$$dp_i = \sum_{j\in V\setminus\{i\}} \tilde{m}_{ji}(p_j - p_i) + \sqrt{\frac{1}{\rho_e} p_i(1-p_i)}dB_i,$$

où $(B_i)_{i\in V}$ sont des mouvements browniens i.i.d, ρ_e est un paramètre contrôlant la vitesse de diffusion, et où $(\tilde{m}_{ij})_{i,j\in V}$ correspond aux taux de migrations d'un patch à l'autre. Afin de faire en sorte que le nombre d'individus dans chaque patch reste constant, $(\tilde{m}_{ij})_{i,j\in V}$ vérifie de plus

$$\forall i \in V, \sum_{j\in V\setminus\{i\}} \tilde{m}_{ij} = \sum_{j\in V\setminus\{i\}} \tilde{m}_{ji}.$$

Remarque 1.2.9. Les taux de migration $(\tilde{m}_{ij})_{i,j\in V}$ doivent aussi correspondre aux arêtes du graphe $\mathcal{G}$: pour tout $(i, j) \in V$, $\tilde{m}_{i,j} \neq 0$ si, et seulement si (i, j) est une arête du graphe $\mathcal{G}$. Dans la suite, nous allons toutefois supposer que $\mathcal{G}$ n'est pas fixé à l'avance, mais se déduit des taux de migration.

Il est possible d'appliquer cette idée au modèle de Wright-Fisher de la façon suivante. Supposons que pour tout $i \in V$, le patch i contient N_i individus, et soit $(m_{ij})_{i,j\in V}$ telle que

$$\begin{aligned} &\forall (i, j) \in V^2, m_{ij} \geq 0, \\ &\text{et } \forall i \in V, \sum_{j\in V\setminus\{i\}} m_{ij} = \sum_{j\in V\setminus\{i\}} m_{ji} = 1 - m_{i,i}. \end{aligned}$$

La dynamique est alors définie de la façon suivante. A chaque génération et dans chaque patch $i \in V$, chaque nouvel individu choisit son parent dans le même patch avec probabilité m_{ii}, ou dans le patch $j \neq i$ avec probabilité m_{ji}. Comme dans le modèle de Wright-Fisher initial, le parent est choisi uniformément au hasard parmi tous les individus du patch, indépendamment de leurs types.

Remarque 1.2.10. Une façon alternative de définir un modèle de Wright-Fisher avec une structuration spatiale de type stepping-stone consiste à fixer à l'avance les nombres M_{ij}, $i \geq j$ d'individus

migrant du patch i au patch j. Afin de garantir des effectifs constants par patch, ces nombres doivent vérifier

$$\forall i \in V, \sum_{j \in V \setminus \{i\}} M_{ij} = \sum_{j \in V \setminus \{i\}} M_{ji} \quad \leq N_i.$$

Que la structuration spatiale soit de type *modèle d'îles* ou de type *stepping-stone*, il est dans les deux cas possible de construire un processus dual encodant les généalogies. Pour cela, supposons que les patchs sont indexés par un ensemble V ($V = [\![1, I]\!]$ dans le cas du modèle d'îles). Comme dans le cas du modèle de Wright-Fisher sans structuration spatiale, nous cherchons à reconstruire les généalogies d'un échantillon de m individus. Cependant, cette fois-ci, pour que deux individus soient des descendants du même parent, il est nécessaire que leurs parents proviennent du même patch. Il faut donc introduire de la structuration spatiale dans le processus dual, ce qui conduit à construire un coalescent structuré (voir par exemple [Not90; Not97; Wil98]). En fonction de la question d'intérêt, ce processus peut être défini sur un espace d'états différent :

- Si l'on s'intéresse au nombre d'ancêtres d'un échantillon, alors le m-coalescent structuré peut être défini sur l'espace d'états

$$\mathcal{S}_1^{(m)} := \left\{ (n_i)_{i \in V} : \sum_{i \in V} n_i \leq m \right\}.$$

Ceci correspond à ne suivre que le nombre d'ancêtres présents dans chacun des patchs. L'intérêt est que le processus associé est plus simple à étudier, mais de l'information sur les généalogies est perdue.

- Si l'on cherche à reconstituer les généalogies complètes de l'échantillon, alors il faut plutôt travailler sur l'espace des partitions marquées

$$\mathcal{S}_2^{(m)} := \left\{ (\pi^{(1)}, i_1), ..., (\pi^{(l)}, i_l) : \left\{ \pi^{(1)}, ..., \pi^{(l)} \right\} \in \mathcal{P}_m \text{ et } \forall l' \in [\![1, l]\!], i_{l'} \in V \right\}.$$

Comme précédemment, chaque bloc de la partition correspond à des individus ayant trouvé un ancêtre commun. Cette fois-ci, les blocs sont de plus "marqués" par la localisation de l'ancêtre correspondant.

Afin d'illustrer ce concept, considérons le cas du modèle de Wright-Fisher avec structuration spatiale de type stepping-stone introduit plus haut. Considérons m individus, numérotés 1,..., m et pris dans les patchs $i_1, ..., i_m \in V$. Alors :

- L'individu d'indice $l \in [\![1, m]\!]$ provient d'un parent situé dans le patch j avec probabilité m_{j,i_l}.
- Si deux individus proviennent de parents appartenant au même patch j, alors ce parent est le même avec probabilité $1/N_j$.

Remarque 1.2.11. Le modèle de Wright-Fisher avec structuration spatiale de type stepping-stone permet de comprendre dans quelle mesure la structuration spatiale modifie les généalogies. En effet, nous avons vu précédemment que dans le modèle de Wright-Fisher simple, deux individus vont toujours finir par trouver un ancêtre commun en un temps fini. Avec la structuration spatiale, cela dépend des propriétés des marches aléatoires associées aux lignées ancestrales. Ainsi, si le nombre d'individus par patch est borné :

- Si les lignées ancestrales sont (presque sûrement) une infinité de fois dans le même patch, alors elles finissent presque sûrement par fusionner. Les individus correspondants descendent alors d'un même ancêtre, et ont le même type.
- Sinon, elles ne fusionnent jamais avec une probabilité non nulle. Il est alors possible d'avoir coexistence de plusieurs types pour toujours.

Voir la Section 6.3 de [Eth11] pour une discussion dans le cas du modèle d'îles.

Populations de tailles non constantes

Les versions du modèle de Wright-Fisher que nous avions considérées jusque là supposent que la population est de taille constante (ou localement constante). Cette hypothèse permet de remonter dans le temps et de reconstruire les généalogies d'un échantillon d'individus. En effet, connaître la taille de la population permet de calculer les probabilités que plusieurs individus descendent d'un même parent dans la génération précédente.

De ce fait, si les fluctuations de tailles de populations sont connues à l'avance, il est alors possible de construire une variante du modèle de Wright-Fisher dans laquelle la population est de taille non constante, sans perdre le processus dual associé. Les fluctuations de taille de population peuvent être déterministes (voir par exemple [CW10; DT95; Shp+10; SH91]), ou encore données par une chaîne de Markov [KK03; SSI04]. Dans les deux cas, ceci implique de définir à l'avance quelle dynamique suit l'évolution du nombre d'individus. De plus, pour remonter dans le temps, cette dynamique doit pouvoir être réversible.

S'il est possible d'appliquer cette approche au cas de populations en expansion (par exemple, [SH91] traite le cas d'une croissance exponentielle déterministe), ceci impose cependant de fixer la dynamique suivie par l'expansion, plutôt qu'avoir un modèle générant une expansion comme conséquence de la dynamique de reproduction. Par conséquent, cette approche ne permet pas d'étudier certaines propriétés d'une expansion, telles que sa vitesse (celle-ci est fixée à l'avance) ou les jeux de paramètres pour lesquels l'expansion est effectivement possible.

Dans le cas de populations affectées par des extinctions locales fréquentes, un autre moyen d'intégrer les extinctions peut aussi être utilisé. Basé sur un modèle d'îles, il a été initialement introduit dans [Sla77] et généralisé dans [WM90]. Dans ce modèle, en plus de la dynamique de reproduction, les patchs peuvent être affectés par des événements d'extinction, qui tuent tous les individus dans le patch. Tout patch éteint est toutefois aussitôt recolonisé par les descendants d'un nombre fixé d'individus, qui peut être potentiellement beaucoup plus petit que la taille du patch. Selon la variante considérée, ces individus viennent tous du même patch ou bien de patchs différents (ou une combinaison des deux, voir [WM90]).

Les patchs vides étant aussitôt recolonisés, cette approche alternative n'est toujours pas adaptée à l'étude de potentielles extinctions globales, elle permet toutefois d'illustrer l'intérêt de s'intéresser aux populations vivant dans un environnement fragmenté et perturbé du point de vue de la génétique des populations. En effet, si les généalogies ont la même forme (à changement d'échelle de temps près) qu'en l'absence de structuration spatiale et d'événements d'extinction lorsque ceux-ci affectent chaque patch indépendamment [Wak98; Wak04; WA01], la situation est très différente lors d'événements d'extinction de masse. Dans [TV09], les auteurs s'intéressent ainsi au cas d'événements d'extinction affectant un grand nombre de patchs d'un coup, qui sont aussitôt recolonisés par les descendants d'un *même* groupe d'individus provenant d'un *même* patch source. Un petit nombre d'individus donne donc naissance à un très grand nombre de descendants, et dans la limite, il est possible d'observer des fusions de plus de deux lignées ancestrales dans le processus dual qui encode les généalogies. Le coalescent obtenu comporte ainsi des fusions multiples, et est appelé Ξ-coalescent.

Intégration d'une banque de graines

Nous allons maintenant nous intéresser à une autre extension du modèle de Wright-Fisher, qui permet de prendre en compte la capacité d'une espèce à former une banque de graines ou à rester dormante pendant un certain temps. Ici, je désignerai par "capacité à former une banque de graines" ou "capacité à entrer en dormance" (sous forme de graines dans le cas de plantes) le fait que certains individus peuvent rester dormants pendant au moins une génération complète sans perdre en viabilité. Je ne m'intéresserai donc pas au cas des *banques de graines temporaires*

("transient seed banks" en anglais), dans lesquelles les graines ne restent dormantes que durant quelques mois, par exemple le temps que le printemps arrive. Voir [BB14] pour une présentation plus complète des différents types de banque de graines qui peuvent être considérés.

Le premier modèle de Wright-Fisher avec banque de graines a été introduit dans [KKL01]. Dans ce modèle, les graines peuvent être dormantes durant $H \geq 0$ générations complètes, le cas $H = 0$ correspondant à l'absence de banque de graines. La banque de graines n'est toutefois pas modélisée explicitement, mais de la façon suivante : lorsqu'un individu produit durant la génération n choisit un parent, il ne le prend pas forcément dans la génération $n-1$, mais dans l'une des générations $n-1$ $n-2$,..., $n-H-1$, respectivement avec probabilités b_1, ..., b_{H+1}. Une conséquence directe de cette construction est que $(X(n))_{n\geq 0}$ n'est pas markovien. Il est cependant possible de rendre le processus markovien, en le considérant sur une fenêtre de $H+1$ générations. Ainsi, $(X(n), ..., X(n+H))_{n\geq 0}$ est markovien. En utilisant cette transformation, les auteurs de [KKL01] montrent que lorsque le nombre d'individus N tend vers $+\infty$ et en comptant en unités de N générations, les généalogies d'un échantillon convergent vers un coalescent de Kingman ralenti : au lieu de fusionner à taux 1, chaque paire de blocs fusionne à taux β_1^2, où

$$\beta_1 = \frac{1}{\sum_{i=1}^{H+1} i b_i}$$

est l'inverse de l'âge moyen d'une graine lorsqu'elle germe.

Dans [Bla+13], ce modèle est généralisé en considérant que les graines peuvent potentiellement rester viables pendant des durées arbitrairement longues. Dans ce nouveau modèle, lorsqu'un individu produit durant la génération n choisit un parent, il le prend dans la génération $n-i$, où i suit la loi de probabilité μ sur $\mathbb{N}\backslash\{0\}$. Le cas où μ est de support borné correspond au modèle introduit dans [KKL01], mais μ peut maintenant aussi être de support non borné. Dans ce cas, il n'est alors plus possible d'effectuer la transformation précédemment décrite pour obtenir un processus markovien. L'approche utilisée dans [Bla+13] consiste à reformuler le problème en termes de processus de renouvellement.

Si l'âge moyen d'une graine lors de sa germination est borné (i.e, si $\sum_{i=0}^{+\infty} i\mu(\{i\}) < +\infty$), alors les généalogies d'un échantillon convergent là encore vers un coalescent de Kingman ralenti d'un facteur $(\sum_{i=0}^{+\infty} i\mu(\{i\}))^2$. Sinon, il est possible d'observer un comportement plus riche. Pour illustrer ceci, [Bla+13] étudie le cas où il existe $\alpha > 0$ et $L : \mathbb{N}\backslash\{0\} \to \mathbb{R}_+$ à variations lentes tels que

$$\forall n \in \mathbb{N}\backslash\{0\}, \mu(\{n, n+1, ...\}) = n^{-\alpha} L(n).$$

Si $\alpha > 1$, alors $\sum_{i=0}^{+\infty} i\mu(\{i\}) < +\infty$ et nous retrouvons un coalescent de Kingman. Si $1/2 < \alpha < 1$, deux individus trouvent toujours un ancêtre commun, mais en un temps d'espérance infinie. Enfin, si $\alpha < 1/2$, deux lignées ancestrales ne fusionnent jamais avec probabilité non nulle.

Ces résultats indiquent qu'il est nécessaire de changer d'échelle de temps afin d'observer les généalogies complètes d'un échantillon d'individus et pouvoir les comparer à un coalescent de Kingman. Mais pour cela, encore faut-il pouvoir identifier l'échelle de temps en question, et ceci est rendu compliqué par le fait que le modèle est non markovien. Ainsi, dans [Bla+15b; Gon+14], les auteurs s'intéressent au cas où

$$\mu = \mu(N) = (1-\epsilon)\delta_1 + \epsilon\delta_{N^\beta},\ \beta > 0,\ \epsilon \in (0,1),$$

et montrent que si $\beta < 1/3$, alors l'échelle de temps sur laquelle les coalescences peuvent être observées est $N^{1+2\beta}$, et les généalogies d'un échantillon convergent alors vers un coalescent de Kingman. Généraliser ces résultats au cas $\beta \geq 1/3$ est cependant difficile du fait du caractère non markovien du processus.

Le modèle introduit dans [Bla+16] permet de pallier ce problème, en modélisant la banque de graines de façon explicite, et en supposant que chaque graine germe après un temps distribué

selon une loi géométrique. Ceci permet d'obtenir un processus markovien, plus facile à analyser. Le modèle est défini de la façon suivante.

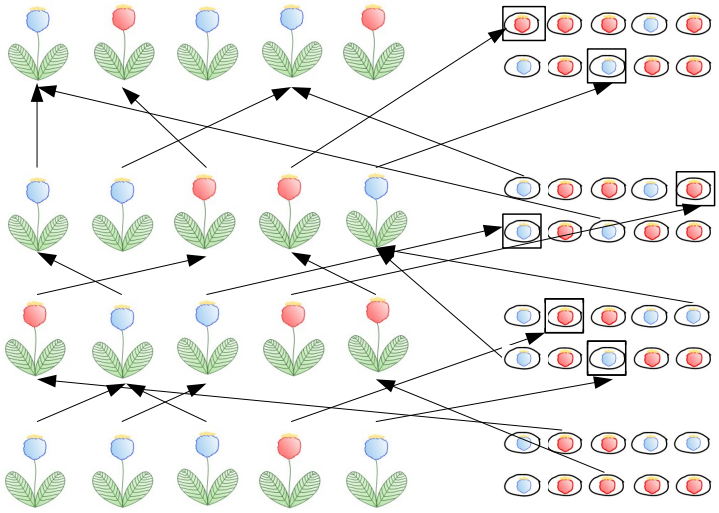


Figure 1.8: Illustration de la dynamique du modèle de Wright-Fisher avec banque de graines de [Bla+16], dans une population de $N = 5$ plantes et $M = 10$ graines. Les plantes et graines de type A (resp. a) sont représentées en bleu (resp. rouge). Ici, $\lfloor \epsilon N \rfloor = 2$.

Définition 1.2.12. *(Modèle de Wright-Fisher avec banque de graines [Bla+16]) Soient $N, M \in \mathbb{N}\backslash\{0\}$, et soit $\epsilon \in [0,1]$ tel que $\epsilon N \leq M$. Soient de plus $N_0, M_0 \in \mathbb{N}\backslash\{0\}$. Le modèle de Wright-Fisher avec banque de graines $(X(n), G(n))_{n\in\mathbb{N}}$ de condition initiale (N_0, M_0) est la chaîne de Markov à valeurs dans $[\![0, N]\!] \times [\![0, M]\!]$ décrivant l'évolution du nombre $X(n)$ de plantes de type a et du nombre $G(n)$ de graines de type a dans une population de N plantes et de M graines pouvant être de type a ou A, et tel que $(X(0), G(0)) = (N_0, M_0)$. La dynamique du modèle est de plus décrite de la façon suivante.*

- *A la génération $n+1$, $N - \lfloor \epsilon N \rfloor$ plantes sont des descendants directs des plantes présentes durant la génération n. Chacune choisit ainsi un parent uniformément au hasard parmi les N plantes de la génération n.*
- *Les $\lfloor \epsilon N \rfloor$ plantes restantes sont issues de graines de la banque de graines qui germent, et donnent leur type aux plantes correspondantes.*
- *Afin de maintenir un nombre constant de graines dans la banque de graines, $\lfloor \epsilon N \rfloor$ nouvelles graines intègrent la banque de graines, chacune provenant d'une plante choisie uniformément au hasard parmi les N plantes de la génération n.*

Voir la Figure 1.8 pour une illustration de la dynamique.
Pour comprendre comment définir un processus dual, observons que les individus se choisissent :

- ou bien un parent "plante", présent dans la génération précédente,
- ou bien un parent "graine", lui-même descendant d'une plante produite plusieurs générations en arrière dans le temps.

Le processus dual associé à ce modèle ressemble donc au coalescent structuré : chaque bloc est marqué comme correspondant à une *plante* ou à une *graine*. De même, le modèle forwards-in-time peut être interprété un modèle d'îles pour lequel $I = 2$. Les "migrations" correspondent ici aux événements de germination et de production de graines, à ceci près que contrairement au modèle d'îles, la reproduction est gelée dans l'"île" contenant les graines. Voir la Figure 1.9 pour une illustration de la relation de dualité.

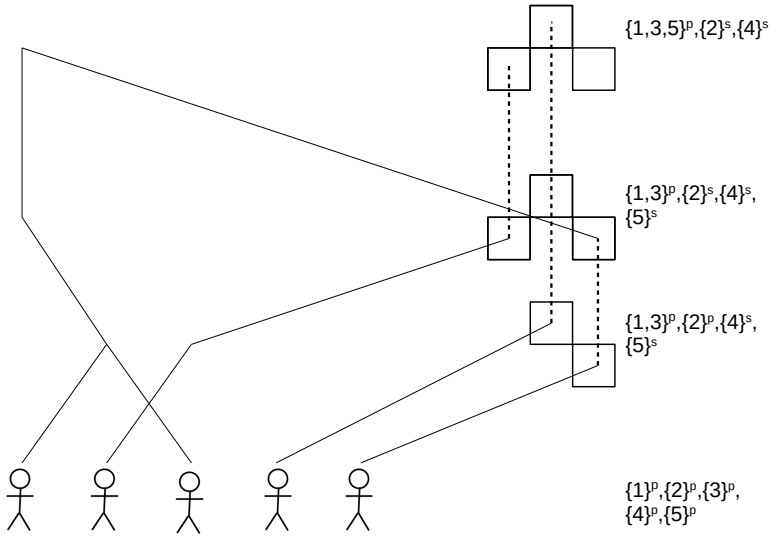


Figure 1.9: Généalogies d'un échantillon dans un modèle de Wright-Fisher avec banque de graines. Les carrés représentent des emplacements dans la banque de graines. La dynamique *forwards-in-time* correspond à celle illustrée par la Figure 1.8.

Dans [Bla+16], les auteurs s'intéressent au régime de paramètres sous lequel la fraction ϵ de la banque de graines qui germe à chaque génération est proportionnelle à $\epsilon \propto N^{-1}$. Ainsi, le nombre de générations durant lesquelles une graine reste dormante suit une loi géométrique de paramètre $\propto N^{-1}$. Plus précisément, les auteurs supposent l'existence de $c, K \in (0, \infty)$ tels que $\epsilon = cN^{-1}$ et $M = NK^{-1}$. En faisant tendre $N \to +\infty$ et en comptant en unités de N générations, le modèle de Wright-Fisher avec banque de graines converge vers la *diffusion de Wright-Fisher avec banque de graines*, définie de la façon suivante.

Définition 1.2.13. *Soient $(p^0, s^0) \in [0,1]^2$. La diffusion de Wright-Fisher avec banque de graines $(p_t, s_t)_{t\geq 0}$ de condition initiale (p^0, s^0) et de paramètres $c, K \in (0, \infty)$ est l'unique solution de l'équation différentielle stochastique*

$$\begin{cases} dp_t & = c(s_t - p_t)dt + \sqrt{p_t(1-p_t)}dB_t \\ ds_t & = cK(p_t - s_t)dt \\ (p_0, s_0) & = (p^0, s^0), \end{cases}$$

où $(B_t)_{t\geq 0}$ est un mouvement brownien standard.

Là encore, le processus limite fait penser au modèle *stepping-stone* de Kimura (voir la Section 1.2.2 et [Kim53]), à ceci près que comme la reproduction est gelée dans la banque de graines, il n'y a pas de dérive génétique à l'intérieur.

Le processus limite est associé à un processus dual, appelé *seed bank coalescent*, qui ne correspond pas à un coalescent de Kingman à un changement d'échelle de temps près. Ce n'est

cette fois-ci pas lié à la présence de coalescences multiples comme pour le Ξ-coalescent, mais à la non-satisfaction d'une propriété spécifique appelée "descente de l'infini" (voir [Pit99; Sch00] et la Section 4.1 de [Bla+16]). En effet, dans le coalescent de Kingman, le nombre d'ancêtres d'un échantillon infini d'individus devient instantanément fini. Dans le *seed bank coalescent*, les lignées ancestrales qui restent dans l'"île" correspondant aux plantes coalescent instantanément en un nombre fini de lignées, mais une fraction non nulle des lignées part dans la banque de graines, dans laquelle les coalescences sont bloquées.

Les modèles de [Bla+16] et [KKL01] peuvent ainsi être considérés comme les deux modèles de banque de graines de base en génétique des populations, selon si les graines ne peuvent rester viables que sur de courtes durées (comparé aux temps de coalescence) ou non. Ainsi, le modèle de [KKL01] est plus adapté aux plantes, et le modèle de [Bla+16] aux bactéries. Ces deux modèles peuvent ensuite être enrichis pour prendre en compte des fluctuations des tailles de population (déterministes dans le cas de [ŽT12]), de la sélection naturelle [Koo+17] ou de la structuration spatiale ([HP17], voir aussi [HN21; GHO22] pour une variante à temps continu basée sur le modèle de Moran). De plus, ils peuvent servir de base à des estimateurs permettant de détecter la présence d'une banque de graines [Bla+20] et d'inférer les paramètres associés [Sel+20; Tel+11].

Ajout de la sélection naturelle

Nous nous intéressons maintenant à une autre extension du modèle de Wright-Fisher, qui prend en compte la sélection naturelle. En effet, nous avons supposé jusque là que les deux types a et A étaient neutres : un individu de type a a ainsi en moyenne autant de descendants qu'un individu de type A. Cependant, comme nous l'avons vu dans la partie consacrée aux motivations biologiques derrière mes travaux de thèse, la diversité génétique n'est pas forcément neutre, et certains allèles peuvent conférer un (dés)avantage sélectif.

Dans toute la suite, nous supposerons que l'allèle A donne un avantage sélectif par rapport à l'allèle a, et que les individus qui le portent ont en moyenne plus de descendants. Une première façon d'intégrer de la sélection naturelle au modèle de Wright-Fisher consiste à changer la façon dont les parents sont choisis. Ainsi, étant donné le nombre $X(n)$ d'individus de type a durant la génération n, chaque descendant choisit maintenant l'un des $X(n)$ parents de type a avec probabilité

$$\frac{X(n)}{(N-X(n))(1+s)+X(n)},$$

ou l'un des $N-X(n)$ parents de type A avec probabilité

$$\frac{(N-X(n))(1+s)}{(N-X(n))(1+s)+X(n)}.$$

Le paramètre s est appelé *coefficient de sélection*, et mesure l'avantage sélectif (ou *fitness relative*) des individus de type A sur les individus de type a.

Sous ces hypothèses, nous avons

$$\begin{aligned}
\forall n \in \mathbb{N}, \mathbb{E}[X(n+1)|X(n)] &= N\frac{X(n)}{(N-X(n))(1+s)+X(n)} \\
&= N\frac{X(n)}{N+(N-X(n))s} \\
&= X(n)\left(\frac{1}{1+\left(1-\frac{X(n)}{N}\right)s}\right) \\
&< X(n).
\end{aligned}$$

Le nombre d'individus tend donc bien à diminuer.

La forme de la formule suggère fortement qu'il va être difficile d'utiliser une approche *forwards-in-time* pour étudier ce modèle. De plus, le choix des parents dépendant des types de *l'ensemble des individus* dans la génération précédente, nous n'avons a priori plus de processus dual à disposition pour reconstituer les généalogies d'un échantillon d'individus. Il va donc être nécessaire d'adapter cette première version du modèle de Wright-Fisher avec sélection afin de retrouver un processus dual associé. Pour cela, intéressons-nous au cas où la sélection est faible, et supposons $s << 1$. La probabilité de choisir un individu de type a comme parent devient alors

$$\begin{aligned}\mathbb{P}(\{\text{le parent choisi est de type a}\}) &= \frac{X(n)}{N}\left(\frac{1}{1+\left(1-\frac{X(n)}{N}\right)s}\right)\\ &\simeq \frac{X(n)}{N}\left(1-s\left(1-\frac{X(n)}{N}\right)\right)\\ &\simeq (1-s)\frac{X(n)}{N}+s\left(\frac{X(n)}{N}\right)^2.\end{aligned}$$

Cette approximation peut être interprétée de la façon suivante :

- Avec probabilité $1-s$, un seul parent est choisi, et l'individu produit prend son type, comme en l'absence de sélection. Nous parlerons d'*événement de reproduction neutre*.
- Avec probabilité s, deux *parents potentiels* sont choisis, et l'individu ne prend le type a que si les *deux parents potentiels choisis* sont de type a, ce qui se produit avec probabilité $X(n)^2N^{-2}$. Nous parlerons cette fois-ci d'*événement de reproduction sélectif.*

Nous supposerons de plus que lors d'un événement sélectif, l'individu descend du premier parent potentiel de type A choisi si un tel parent existe, et du dernier parent potentiel (de type a) choisi sinon.

Cette interprétation nous permet ainsi de déduire une construction alternative pour le modèle de Wright-Fisher avec sélection, basée cette fois-ci sur le choix de plusieurs parents potentiels. D'un point de vue *backwards-in-time* cependant, savoir lequel des deux parents potentiels est le vrai parent nécessite de connaître les types de chacun d'entre eux. Or, nous ne disposons pas de cette information, et c'est justement pour l'obtenir que nous cherchons à reconstruire la généalogie d'un individu. Nous pouvons contourner ce problème en cherchant à reconstruire la généalogie de *chacun des parents potentiels* jusqu'à la condition initiale, afin d'en déduire leurs types et lequel d'entre eux est le vrai parent. Il est donc bien possible de reconstruire les généalogies d'un échantillon d'individus, en adoptant la stratégie suivante.

1. Cherchons l'ensemble des ancêtres potentiels de chacun des individus de l'échantillon, en remontant jusqu'à la condition initiale.
2. Déduisons-en les types de chacun des ancêtres potentiels dans la génération initiale.
3. Redescendons dans le temps jusqu'à la génération présente, en propageant la connaissance des types le long de l'arbre des ancêtres potentiels et en identifiant le vrai parent associé à chaque événement de reproduction sélectif.

Remarque 1.2.14. A noter que si l'objectif est juste d'étudier la proportion d'individus de type a dans un échantillon, l'étape 3 de propagation des types le long de l'arbre des ancêtres potentiels n'est pas nécessaire. En effet, pour qu'un individu soit de type A, il suffit qu'au moins un de ses ancêtres potentiels au temps 0 soit de type A. Nous pouvons donc conclure dès la fin de l'étape 2.

Lorsque le coefficient de sélection est de la forme $s_N = \alpha N^{-1}$ et sous le changement d'échelle de temps classique (temps accéléré d'un facteur N, et $N \to +\infty$), la proportion $(p_t^A)_{t\geq 0}$ d'individus du type A avantagé sélectivement converge vers la *diffusion de Wright-Fisher avec sélection*, qui est la solution de l'équation différentielle stochastique

$$dp_t^A = \alpha p_t^A(1-p_t^A) + \sqrt{p_t^A(1-p_t^A)}dB_t,$$

où $(B_t)_{t\geq 0}$ est un mouvement brownien standard. Le terme déterministe traduit le fait que la sélection n'a un effet visible que lorsque les deux parents potentiels choisis sont de types différents. De plus, le processus dual associé au modèle de Wright-Fisher avec sélection converge vers l'*Ancestral Selection Graph* [KN97; NK97], qui ressemble au coalescent de Kingman à ceci près qu'en plus des fusions de lignées, chaque lignée peut aussi se diviser en deux à taux α. Ceci permet d'ailleurs de comprendre quelles seraient les différences observées en prenant un scaling différent pour le coefficient de sélection :

- Si $s_N = o(N)$, alors toutes les lignées ancestrales coalescent avant que le premier événement de reproduction sélectif ne se produise. Il y a donc déjà eu fixation de l'un des deux types, et la sélection naturelle ne laisse pas de trace visible.

- Si $s_N N \xrightarrow[N\to+\infty]{} +\infty$, alors dans le changement d'échelle de temps correspondant au coalescent de Kingman, le nombre de lignées d'ancêtres potentiels d'un individu devient instantanément infini. Il faut donc considérer une autre échelle de temps et/ou d'autres objets pour étudier les propriétés du modèle.

Si le choix de *deux* parents potentiels est motivé par la version initialement introduite du modèle de Wright-Fisher avec sélection, il est toutefois possible de considérer des variantes dans lesquelles $k \geq 2$ parents potentiels sont choisis, ou dans lesquelles le nombre de parents potentiels est aléatoire, par exemple pour prendre en compte différentes formes de sélection. De plus, il n'est pas nécessaire de se limiter au cas où $s_N \propto N^{-1}$, et il est possible d'étudier d'autres scalings du coefficient de sélection, voire de considérer le cas $s = 1$. Toutes ces variantes peuvent mener à la construction d'objets mathématiques différents, avec des propriétés caractéristiques. Voir par exemple [Boe+21a; Boe+21b; GS18] (exemple non limités au modèle de Wright-Fisher, et considérant aussi d'autres modèles proches). Dans le cadre de ma thèse, les modèles de populations en expansion que j'ai construits peuvent être interprétés comme des modèles de génétique des populations avec de la sélection forte, pour lesquels $s = 1$ et dans lesquels le nombre de parents potentiels est fixé. Je m'intéresse de plus principalement à la limite $k \to +\infty$, sans changer d'échelle de temps.

1.2.3 Le processus Λ-Fleming Viot spatial

La version spatialisée du modèle de Wright-Fisher vue plus haut est bien adaptée aux populations qui vivent dans un environnement fragmenté, tel que les plantes dans les pieds d'arbres en milieu urbain, chaque patch correspondant alors à un pied d'arbre. Elle peut cependant paraître assez artificielle dans le cas de populations vivant dans un milieu continu, comme les bactéries en croissance dans des boîtes de Pétri, pour lesquelles des modèles de génétique des populations continus en espace seraient plus adaptés. Pourtant, comme expliqué plus haut, définir ce type de modèle n'est en fait pas si facile, du fait du phénomène de "pain in the torus" [Fel75]. Initialement introduit dans [BEV10; Eth08], le processus Λ-Fleming Viot spatial (abrégé dans toute la suite en SLFV) évite ce problème en définissant la dynamique de façon *reproduction-centrée* plutôt qu'*individu-centrée*.

Présentation du processus

Intuitivement, le SLFV peut être vu comme l'évolution de la densité $\omega_t : x \in E \subseteq \mathbb{R}^d \to \omega_t(x) \in [0,1]$ en individus de type a en chaque point de l'espace. Il suppose donc un nombre infini d'individus en chaque point de l'espace. Contrairement au cas de modèles de reproduction individu-centrés, comme le modèle de Wright-Fisher, la dynamique du processus peut être vue comme conduite par un processus ponctuel de Poisson qui encode les événements de reproduction. Ce processus ponctuel de Poisson, que nous noterons Π, est défini sur $\mathbb{R}_+ \times E \times (0,\infty)$. Chaque point $(t,x,\mathcal{R}) \in \Pi$ correspond à un événement de reproduction qui se produit au temps t, et affecte tous les individus à l'intérieur de la boule de centre x et de rayon $\mathcal{R}$ (notée $\mathcal{B}(x,\mathcal{R})$ dans toute la suite). L'intensité du processus ponctuel de Poisson est de $dt \otimes dx \otimes \mu(dr)$, où μ est une mesure σ-finie sur $(0,\infty)$ vérifiant

$$\int_0^\infty \mathcal{R}^d \mu(d\mathcal{R}) < +\infty. \tag{1.2.1}$$

Le SLFV est alors défini de la façon suivante.

Définition informelle 1.2.15. *Soit $\omega^0 : E \to [0,1]$ mesurable. Le processus Λ-Fleming Viot spatial $(\omega_t)_{t\geq 0}$ de condition initiale ω^0 est défini de la façon suivante. Soit Π un processus ponctuel de Poisson défini sur $\mathbb{R}_+ \times E \times (0,\infty)$ d'intensité $dt \otimes dx \otimes \mu(dr)$. Alors, pour tout $(t,x,\mathcal{R}) \in \Pi$, sachant ω_{t-} :*

- *Avec probabilité $(\mathrm{Vol}(\mathcal{B}(x,\mathcal{R})) \cap E)^{-1} \int_{\mathcal{B}(x,\mathcal{R})\cap E} \omega_{t-}(y)dy$, pour tout $y \in \mathcal{B}(x,\mathcal{R}) \cap E$, $\omega_t(y) = 1$.*
- *Sinon, pour tout $y \in \mathcal{B}(x,\mathcal{R}) \cap E$, $\omega_t(y) = 0$.*
- *Indépendamment de la valeur prise par le processus dans $\mathcal{B}(x,\mathcal{R})$, pour tout $y \in E \backslash \mathcal{B}(x,\mathcal{R})$, $\omega_t(y) = \omega_{t-}(y)$.*

De plus, la densité est laissée inchangée en l'absence d'événements de reproduction.

Ceci revient à dire qu'à chaque événement de reproduction $(t,x,\mathcal{R}) \in \Pi$, un individu est choisi uniformément au hasard à l'intérieur de la boule $\mathcal{B}(x,\mathcal{R})$. Il est donc de type a avec probabilité $(\mathrm{Vol}(\mathcal{B}(x,\mathcal{R})))^{-1} \int_{\mathcal{B}(x,\mathcal{R})} \omega_{t-}(y)dy$. Tous les individus dans la boule de reproduction sont alors tués, et remplacés par des descendants de cet individu, qui prennent son type.

Remarque 1.2.16. Dans la version originelle du SLFV [BEV10], les événements de reproduction sont aussi caractérisés par un quatrième paramètre $u \in [0,1]$, appelé *paramètre d'impact* et correspondant à la fraction d'individus remplacés lors de l'événement de reproduction. Ici, nous nous intéresserons uniquement au cas $u = 1$.

Voir la Figure 1.10 pour une illustration de la dynamique. A noter que si le modèle a initialement été défini en supposant que les zones affectées par les événements de reproduction sont des boules, il est en fait tout à fait possible de généraliser la définition à d'autres formes géométriques. Ainsi, dans le Chapitre 3, je m'intéresse plutôt à des ellipses.

Cette définition, bien que très intuitive, n'est en fait pas rigoureuse. En effet, si E n'est pas compact, il se produit un nombre infini d'événements de reproduction à chaque instant, et cette construction ne peut pas être utilisée. Formellement, le SLFV est un processus à valeurs mesures, défini sur l'espace $\widehat{M}_\lambda$ des mesures M sur $E \times \{a,A\}$ dont la marginale sur E est la mesure de Lebesgue, i.e. telles que pour toute fonction $f : E \to \mathbb{R}$ continue et compacte,

$$\int_{\mathbb{R}^d \times \{a,A\}} f(x) M(dx, d\kappa) = \int_{\mathbb{R}^d} f(x)dx.$$

Pour tout $M \in \widetilde{M}_\lambda$, il existe $\omega : E \to [0,1]$ mesurable telle que

$$M(dx, d\kappa) = ((\omega(x)\delta_a(d\kappa) + (1-\omega(x))\delta_A(d\kappa))dx,$$

de sorte que ω peut être interprétée comme une densité en individus de type a. Remarquons que ω n'est pas définie de façon unique, mais à un ensemble de mesure de Lebesgue nulle près. Ceci signifie en particulier qu'il n'est pas possible d'assigner de façon unique une densité en individus de type a à un point de l'espace. Dans cette partie, afin de simplifier les notations, nous utiliserons cependant des densités $(\omega_t)_{t\geq 0}$ pour décrire l'état du SLFV, afin d'alléger les notations.

Le SLFV est défini de façon rigoureuse dans le Chapitre 4, dans une variante avec sélection, mais prendre $k = 1$ dans la définition donnée permet de retrouver le cas sans sélection. Il existe plusieurs façons de définir formellement le processus :

- comme l'unique dual du processus ancestral encodant les généalogies, défini plus loin;
- comme l'unique solution d'un problème martingale;
- en travaillant conditionnellement aux événements de reproduction, et en utilisant le processus dual encodant les généalogies [VW15];
- via une construction de type look-down [EK19; VW15].

Quelle construction utiliser dépend de la question étudiée, chacune étant plus adaptée à des problématiques spécifiques. Ainsi, la construction conditionnellement aux événements de reproduction, qui consiste à reconstituer les généalogies de chaque individu à chaque instant, permet d'associer au processus une densité de façon unique étant donnée une densité initiale. Les constructions de type look-down, quant à elles, font apparaître le SLFV comme la limite en grande densité de modèles individu-centrés. Dans cette thèse, j'utiliserai les deux premières constructions du SLFV.

Même si la définition informelle donnée plus haut n'est pas rigoureuse, elle a toutefois deux intérêts: elle est très visuelle, et elle permet de mettre en valeur le lien avec le processus dual, que je vais maintenant introduire. Afin de simplifier les notations, nous supposerons que $E = \mathbb{R}^d$, mais la définition est généralisable au cas $E \subseteq \mathbb{R}^d$.

De plus, nous nous restreindrons à des conditions initiales appartenant à l'espace $\mathcal{M}_\lambda \subset \widetilde{\mathcal{M}}_\lambda$ des mesures $M \in \widetilde{\mathcal{M}}_\lambda$ telles qu'il existe $\omega : \mathbb{R}^d \to \{0,1\}$ (et non pas $[0,1]$) mesurable satisfaisant

$$M(dx, d\kappa) = ((\omega(x)\delta_a(d\kappa) + (1-\omega(x))\delta_A(d\kappa))dx.$$

Etant donné la dynamique du processus, le SLFV reste alors à valeurs dans $\mathcal{M}_\lambda$ pour tout temps.

Dual et relation de dualité

L'intuition derrière le dual est la suivante : deux individus produits lors d'un même événement de reproduction descendent du même parent, et ont donc le même type. Or, le processus ponctuel de Poisson Π introduit plus haut encode quels événements de reproduction se sont produits, et peut donc être utilisé pour reconstruire les généalogies dans un échantillon d'individus.

Formellement, le processus dual, que nous appellerons le *processus ancestral*, prend ses valeurs dans l'ensemble $\mathcal{M}_p(\mathbb{R}^d)$ des mesures ponctuelles finies sur $\mathbb{R}^d$, et est défini de la façon suivante.

Définition 1.2.17. *(Processus ancestral) Soit $\Xi^0 \in \mathcal{M}_p(\mathbb{R}^d)$. Le processus ancestral $(\Xi_t)_{t\geq 0}$ de condition initiale Ξ^0 est le processus Markovien de saut à valeurs dans $\mathcal{M}_p(\mathbb{R}^d)$ défini de la façon suivante. Posons d'abord $\Xi_0 = \Xi^0$. Puis, soit $\overleftarrow{\Pi}$ un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty)$ d'intensité $dt \otimes dx \otimes \mu(dr)$. Pour tout $(t,x,\mathcal{R}) \in \overleftarrow{\Pi}$ affectant au moins un atome de Ξ_{t-}, sachant $\Xi_{t-} = \sum_{i=1}^{N_{t-}} \delta_{x_i}$:*

- *Nous choisissons $y \in \mathcal{B}(x, \mathcal{R})$ uniformément au hasard dans $\mathcal{B}(x, \mathcal{R})$.*
- *Si $\mathcal{S}^{\mathcal{R}}(\Xi_{t-})$ représente l'ensemble des entiers $i \in [\![1, N_{t-}]\!]$ tels que $x_i \in \mathcal{B}(x, \mathcal{R})$, nous posons alors*

$$\Xi_t = \Xi_{t-} - \sum_{i \in \mathcal{S}^{\mathcal{R}}(\Xi_{t-})} \delta_{x_i} + \delta_y.$$

Intuitivement, ce processus peut être interprété de la façon suivante. La condition initiale Ξ^0 correspond aux positions occupées par les individus dont nous cherchons à reconstruire les généalogies. Si nous remontons dans le temps, à chaque événement de reproduction, les individus affectés, qui viennent d'être produits, sont remplacés par le parent qui leur a donné naissance. Le nombre d'atomes de Ξ_t correspond au nombre d'ancêtres différents pour l'échantillon, et ne peut que diminuer au cours du temps. En particulier, si Ξ_t ne comporte plus qu'un atome, alors tous les individus de l'échantillon descendent d'un même ancêtre commun. Le point important dans la définition est le fait que le temps s'écoule dans le sens inverse par rapport au SLFV : nous remontons ici dans le passé, et cherchons les événements de reproduction qui se sont produits, du plus récent au plus vieux.

Du fait de la condition (1.2.1) satisfaite par μ, le processus ancestral est bien défini. En effet, observons que le nombre d'atomes dans le processus ancestral ne peut que diminuer. De plus, chaque atome $x \in \mathbb{R}^d$ est affecté par un événement de reproduction à taux

$$\int_0^\infty \mathrm{Vol}(\mathcal{B}(x, \mathcal{R}))\mu(d\mathcal{R}) \propto \int_0^\infty \mathcal{R}^d \mu(d\mathcal{R}) \quad < +\infty.$$

Ainsi, le processus ancestral saute à taux borné, et nous pouvons conclure.

Dans [VW15], le processus ancestral est utilisé pour définir le SLFV, en fixant à l'avance à la fois le processus ponctuel de Poisson Π qui encode les événements de reproduction et les positions des parents choisis lors de chaque événement de reproduction. Ceci conduit à travailler avec ce que nous appellerons un processus ponctuel de Poisson étendu, qui encode l'ensemble des caractéristiques des événements de reproduction, à l'exception du type des parents. Ce processus, que nous noterons $\overrightarrow{\Pi}$, est défini sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0, \infty) \times \mathcal{B}(0, 1)$ et d'intensité $dt \otimes dx \otimes \mu(d\mathcal{R}) \otimes \bar{u}(dp)$, où $\bar{u}$ est la loi uniforme sur $\mathcal{B}(0, 1)$. Puis, à chaque instant $t \geq 0$, la portion du processus ponctuel de Poisson étendu correspondant aux événements de reproduction qui se produisent sur l'intervalle $[0, t]$ est utilisée pour construire un nouveau processus ponctuel de Poisson étendu $\overleftarrow{\Pi}_t$. Ce processus encode les mêmes événements de reproduction, mais ordonnés en remontant dans le temps à partir de l'instant t. Le processus $\overleftarrow{\Pi}_t$ peut alors être utilisé pour construire le processus ancestral de condition initiale δ_x sur l'intervalle $[0, t]$, pour tout $x \in \mathbb{R}^d$. La position du processus ancestral au temps t, notée x_0, correspond à l'ancêtre au temps 0 des individus en x au temps t. Cet ancêtre est de type $1 - \omega_0(x_0)$, et nous pouvons ainsi poser $\omega_t(x) = \omega_0(x_0)$. La construction repose sur l'invariance par changement de sens du temps de la distribution du processus ponctuel de Poisson, qui permet de construire directement $\overleftarrow{\Pi}_t$ à partir de $\overrightarrow{\Pi}$.

Introduisons maintenant la relation de dualité satisfaite par le SLFV et le processus ancestral.

Proposition 1.2.18. *Soient $l \in \mathbb{N}^*$ et ψ une fonction intégrable sur $(\mathbb{R}^d)^l$. Si $(\Xi_t)_{t\geq 0} = (\sum_{i=1}^{N_t} \delta_{\xi_t^i})_{t\geq 0}$ est un processus ancestral, alors pour tout $t \geq 0$,*

$$\mathbb{E}_{\omega_0 = \omega^0} \left[\int_{(\mathbb{R}^d)^l} \psi(x_1, ..., x_l) \left\{ \prod_{j=1}^{l} \omega_t(x_j) \right\} dx_1 ... dx_l \right]$$
$$= \int_{(\mathbb{R}^d)^l} \psi(x_1, ..., x_l) \mathbb{E}_{\Xi_0 = \Xi[x_1, ..., x_l]} \left[\prod_{i=1}^{N_t} \omega_0(\xi_t^i) \right] dx_1 ... dx_l.$$

Intuitivement, cette proposition signifie que la probabilité que l individus situés en $x_1, ..., x_l$ au temps t soient tous de type a est égale à la probabilité que tous leurs ancêtres au temps 0 soient de type a. L'intégration contre la fonction ψ permet à la formule de dualité de ne pas dépendre du choix de la densité ω_t associée au processus à valeurs mesures. Voir la Proposition $3.1.8$ du Chapitre 3 pour une formulation plus rigoureuse de la relation de dualité (le cas $k = 1$ correspondant au SLFV classique défini plus haut).

Ajout de la sélection naturelle

De la même façon que pour le modèle de Wright-Fisher, il est possible de modifier le SLFV de façon à y incorporer d'autres éléments. Dans le cas du SLFV, beaucoup d'articles se sont intéressés à l'ajout de différentes formes de sélection naturelle [BEK21; CK19; EFP17; Eth+17; EFS17; EVY20; FP17; KR20]. La première version du SLFV avec sélection a été introduite dans [EVY20]. Dans ce modèle, aux événements de reproduction définis précédemment, correspondant à des événements *neutres*, s'ajoutent des événements *sélectifs*. La différence avec les événements neutres est la suivante : au lieu de choisir un seul parent potentiel, deux *parents potentiels* sont échantillonnés. En supposant que le type avantagé sélectivement est le type A, le vrai parent de l'événement de reproduction est alors le premier parent potentiel de type A choisi si un tel parent potentiel existe, et le dernier parent potentiel de type a choisi sinon. Le processus dual obtenu ressemble à l'*Ancestral Selection Graph* mentionné plus haut, mais dans une version structurée en espace, et définie sur un espace continu.

Le SLFV avec sélection et son dual ont été étudiés sous deux régimes de paramètres différents : l'un dans lequel le paramètre d'impact u tend vers 0 [EVY20; FP17], et l'autre dans lequel le paramètre d'impact u est constant égal à 1 [Eth+17; EFS17]. Dans les deux cas, la fréquence des événements de sélection comparée aux événements neutres tend elle aussi vers 0. Ceci correspond à un régime de sélection faible, et il est nécessaire de changer d'échelle de temps pour observer l'effet de la sélection.

Des mécanismes plus complexes de sélection ont aussi été étudiés. Ainsi, [FP17] s'intéresse à une forme très générale de sélection, [EFP17] étudie le cas de la sélection contre l'hétérozygotie dans des populations diploïdes, et [BEK21; CK19; KR20] traitent du cas d'une sélection fluctuante (ou de façon équivalente, d'un environnement aléatoire).

La preuve du caractère bien défini de ces différents modèles repose sur la preuve issue de [EVY20], traitant du cas de la forme de sélection la plus simple, mais pouvant être généralisée à d'autres formes de sélection. Cette preuve peut servir de base pour définir d'autres variantes du SLFV comme uniques solutions d'un problème martingale. La structure de la preuve est la suivante :

- *Etape 1*: Montrer l'existence d'une solution au problème martingale.

  Ceci passe par l'utilisation de la définition intuitive du processus à partir d'un processus ponctuel de Poisson. En effet, si cette construction n'est pas valable sur $\mathbb{R}^d$, elle est en revanche valide sur tout compact $E \subset \mathbb{R}^d$. Nous pouvons donc considérer une suite $(E_n)_{n\in\mathbb{N}}$ de compacts de $\mathbb{R}^d$ croissante pour l'inclusion tels que $E_n \xrightarrow[n\to+\infty]{} \mathbb{R}^d$. Nous obtenons une suite $((M_t^n)_{t\geq 0})_{n\in\mathbb{N}}$ de SLFVs avec sélection, chacun défini sur E_n via la construction utilisant le processus ponctuel de Poisson, qui est maintenant bien définie de par la condition (1.2.1). Il suffit pour conclure de vérifier que la limite est bien solution du problème martingale conjecturé comme caractérisant le processus.

- *Etape 2*: Construire un processus dual candidat.

  L'idée de la construction est la même que pour le processus ancestral dual du SLFV classique, sauf que cette fois-ci, comme les deux parents potentiels doivent être conservés, le

nombre d'atomes n'est pas forcément décroissant. Il est toutefois possible de borner le nombre d'atomes par le nombre d'individus dans un processus de Yule, ce qui permet de contrôler le taux de saut du processus et de conclure.

- *Etape 3*: Vérifier que le processus dual candidat est bien le processus dual.

  La preuve repose sur l'utilisation des problèmes martingales associés au SLFV et à son dual (candidat pour l'instant). La première étape consiste à élargir le problème martingale à une plus grande famille de fonctions test, et la deuxième est une adaptation de la preuve du Théorème 4.4.11 dans [EK86].

Dans le Chapitre 3, je me base sur ce schéma de preuve pour construire une variante du SLFV adaptée aux populations en expansion. Ceci nécessite toutefois quelques adaptations. Ainsi, l'existence d'une solution au problème martingale est obtenue via un argument de couplage, et contrôler le taux de saut du processus dual nécessite de considérer une condition plus stricte que (1.2.1).

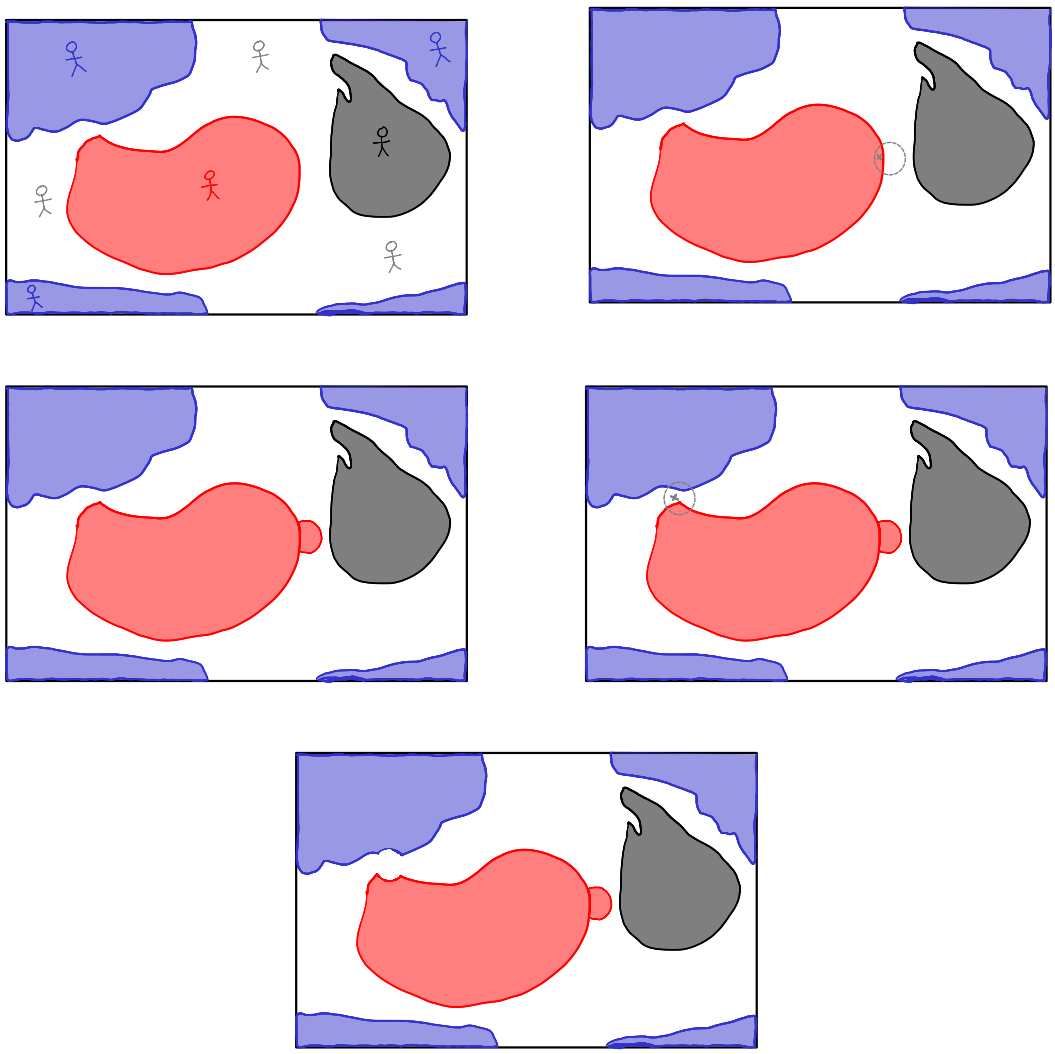


Figure 1.10: Définition heuristique du SLFV. (a) Ici, nous supposons qu'initialement, les individus occupant une même position de l'espace sont du même type. (b) Nous attendons qu'un premier événement de reproduction affecte la zone. Un parent est alors choisi uniformément au hasard dans la zone. (c) L'individu se reproduit, et ses descendants remplissent la boule correspondant à la zone affectée, tandis que les autres individus dans la zone meurent. (d), (e) Nous répétons ceci à chaque nouvel événement de reproduction affectant la zone.

1.3 Modèles de génétique des populations pour les populations en expansion

De la même façon que la plupart des autres modèles de génétique des populations, le modèle de Wright-Fisher et le processus Λ-Fleming Viot spatial supposent que la population considérée est de taille constante. Comme expliqué plus haut, cette hypothèse est nécessaire pour disposer d'un processus dual encodant les généalogies. Il est de plus possible de relaxer cette condition dans le cas du modèle de Wright-Fisher, à condition de définir à l'avance le scénario d'expansion ou de fluctuations. La dynamique de l'expansion est alors fixée, et les fluctuations stochastiques sont prédéfinies plutôt que directement générées par l'aléa dans la reproduction des individus. Nous avons donc besoin d'une autre façon de modéliser expansions et fluctuations de taille de populations, dans laquelle les variations d'effectif seraient une conséquence de la dynamique de reproduction des individus, sans perdre pour autant le processus dual.

Dans cette section, je commencerai par présenter l'approche introduite dans [DF16; HN08]. Issue de la théorie de systèmes de particules en interaction, elle a permis d'obtenir de premiers résultats théoriques sur la diversité génétique au front de populations en expansion. Parallèlement, j'exposerai comment étudier la diversité génétique dans ce type de modèle, et je mettrai en valeur dans quelle mesure les résultats et techniques présentés peuvent être généralisés à d'autres modèles. Puis, j'expliquerai comment j'ai appliqué cette idée au processus Λ-Fleming Viot spatial et à une variante avec banque de graines du modèle de Wright-Fisher.

1.3.1 Individus fantômes

Dans [DF16] et [HN08], les auteurs considèrent des dèmes situés le long d'une ligne, et indexés par $L_n^{-1}\mathbb{Z}$, $L_n > 0$. Chaque dème contient exactement $M_n \in \mathbb{N}\backslash\{0\}$ individus, qui peuvent être de type 1 ou de type 0. D'un point de vue modélisation, les individus de type 0 sont des individus "fantômes", qui correspondent en fait à des emplacements vides, tandis que les individus de type 1 correspondent à des individus réels et observables.

Dans le modèle de [DF16], qui est une variante de celui initialement introduit dans [HN08], la dynamique de reproduction est basée sur un modèle issu de la théorie des systèmes de particules en interaction : le modèle du votant biaisé (ici en faveur des individus de type 1). Dans le modèle du votant, lors d'un événement de vote de la paire ordonnée $(x, y) \in (L_n^{-1}\mathbb{Z} \times [\![1, M_n]\!])^2$, l'individu occupant l'emplacement y "convainc" l'individu en x, qui prend son type. Le biais en faveur des individus de type 1 se traduit par le fait que les individus de type 1 convainquent les autres individus à un taux plus élevé que les individus de type 0. Dans le contexte de la modélisation de populations en expansion, les événements de vote sont interprétés comme des événements de reproduction de la façon suivante : si l'événement implique la paire (x, y), alors l'individu situé en x meurt et est remplacé par un descendant de l'individu en y. C'est d'ailleurs la même idée que dans le modèle de Moran présenté rapidement plus haut (voir Section 1.2.1).

La dynamique du modèle est définie de la façon suivante. Tout d'abord, notons $\xi_t^n(x)$ le type de la cellule en x au temps t. A chaque paire ordonnée (x, y) d'individus pouvant interagir (ici, situés dans des patchs voisins), nous associons deux processus ponctuels de Poisson i.i.d sur $\mathbb{R}_+$: un processus $P^{n,(x,y)}$ d'intensité $r_n > 0$, et un processus $\widetilde{P}^{n,(x,y)}$ d'intensité θR_n^{-1}. Puis,

- A chaque saut de $P^{n,(x,y)}$, l'individu en x meurt, et est remplacé par un descendant de l'individu en y. En d'autres termes, nous posons

$$\xi_t^n(x) = \xi_{t-}^n(y).$$

- A chaque saut de $\widetilde{P}^{n,(x,y)}$, il se passe la même chose, mais *uniquement si* y *est de type* 1. Ceci correspond à poser

$$\xi_t^n(x) = \xi_{t-}^n(y) + (1 - \xi_{t-}^n(y))\xi_{t-}^n(x).$$

Afin de rajouter de la diversité génétique, il est possible d'utiliser des techniques dites de *traceurs* [DF16; HN08]. Plutôt que de rajouter différents types et suivre la densité de chacun, elle consiste à "marquer" certains individus de type 1, et à suivre l'évolution du marquage, qui se transmet aux descendants. Nous introduisons donc la notation $\eta_t^n(x)$ pour encoder le potentiel marquage porté par la cellule en x au temps t. Ainsi, $\eta_t^n(x) = 1$ si cette cellule est de type 1 et marquée, et $\eta_t^n(x) = 0$ sinon.

Pour construire le processus dual, il est possible d'utiliser la même idée que pour le dual du modèle de Wright-Fisher ou du SLFV avec sélection. En effet, à chaque événement de reproduction affectant la paire (x, y) donné par le processus ponctuel de Poisson $\widetilde{P}^{n,(x,y)}$, l'individu en x ne prend le type de l'individu en y au temps $t-$ que si celui-ci est de type 1, et garde le même type sinon. Ainsi, l'individu en x au temps t est de type 0 si, et seulement si les individus en x et y au temps $t-$ sont de type 0. Il est donc possible de considérer les cellules en x et y au temps $t-$ comme les deux "parents potentiels" de l'individu en x au temps t. C'est un peu un abus de langage, car si l'individu en y est de type 0, alors il n'y a pas reproduction, et les individus en x aux temps $t-$ et t sont en fait les mêmes. J'utiliserai cependant cette terminologie afin d'établir une analogie avec les modèles de génétique des populations avec sélection introduits plus hauts.

Malgré ces similarités, du fait des individus fantômes et de l'utilisation de traceurs, le modèle de [DF16] présente une différence fondamentale avec les modèles classiques de génétique des populations avec sélection, liée à la façon dont le type d'un individu peut être déterminé à partir de la connaissance de sa généalogie. Par exemple, dans le modèle de Wright-Fisher avec sélection, le processus dual encodant les ancêtres potentiels permet de déduire le type d'un individu de la façon suivante :

- Reconstruisons la généalogie de cet individu en remontant jusqu'à la condition initiale, et cherchons ses ancêtres potentiels.
- L'individu dont nous cherchons les ancêtres est du type avantagé sélectivement si, et seulement si au moins l'un de ses ancêtres potentiels au temps 0 est du type avantagé.
- Sinon, il est du type désavantagé sélectivement.

Il n'y a donc en particulier pas besoin de savoir quel est le vrai parent parmi tous les ancêtres potentiels pour en déduire le type de l'individu.

Ici, au contraire, nous pouvons distinguer deux niveaux de "diversité génétique" :

- Une "diversité génétique" artificielle, qui correspond aux types 1/réels et 0/fantômes.
- Une diversité génétique observable, qui correspond aux individus de type 1 marqués ou non.

Cette fois-ci, pour connaître le type complet d'un individu (c'est à dire, en incluant un éventuel marquage), il ne suffit pas de connaître les types de tous ses ancêtres potentiels au temps 0 : il faut aussi savoir lequel d'entre eux est le vrai. Ce sera aussi le cas dans les variantes du processus Λ-Fleming Viot spatial et du modèle de Wright-Fisher que j'introduirai plus loin.

Pour répondre à cette question, les auteurs de [DF16] considèrent un dual basé sur une représentation graphique du modèle du votant biaisé [Gri79; Har78]. Ici, je vais présenter un autre type de processus dual, basé sur l'*ordered ASG* utilisé entre autres dans [Len+15] et présenté plus haut, qui consiste à ordonner les différents ancêtres potentiels. L'intérêt de ce processus dual est que sa

construction peut être généralisée à d'autres modèles de populations en expansion, tels que ceux que j'étudierai dans la suite.

Pour illustrer cette approche, intéressons-nous à l'individu en x au temps t, et supposons que le dernier événement de reproduction l'ayant affecté impliquait l'individu en y. S'il s'agissait d'un événement donné par $P^{n,(x,y)}$, alors nous savons que le parent de l'individu est y. Supposons cependant qu'il s'agissait d'un événement donné par $\widetilde{P}^{n,(x,y)}$. L'individu a alors deux "parents potentiels" : y et x. De plus, pour connaître son type, il faut d'abord regarder le type de la cellule en y, puis le type de la cellule en x. Nous ordonnons les deux parents potentiels afin de garder trace de cet ordre. Intéressons-nous ensuite au prochain événement de reproduction à affecter l'un des deux ancêtres potentiels, et supposons là encore qu'il s'agit d'un événement "sélectif".

- Si l'événement affecte la paire (y,z), $z \neq x$, alors nous mettons à jour la liste ordonnée d'ancêtres potentiels, et obtenons : z, y, x.
- Si l'événement affecte la paire (x,z), $z \neq y$, nous obtenons : y, z, x.
- Si l'événement affecte la paire (x,y), la liste ordonnée ne change pas.
- Si l'événement affecte la paire (y,x), nous obtenons cette fois-ci : x, y.

Ce dernier cas permet d'ailleurs d'illustrer le fait que le dernier ancêtre potentiel dans la séquence n'est pas forcément le vrai ancêtre si tous les ancêtres potentiels sont de type 0.

Cet argument est ensuite répété jusqu'à atteindre la condition initiale. Il permet bien de définir un processus dual, car à chaque événement sélectif, le nombre d'ancêtres potentiels augmente d'au plus un. Ainsi, le taux de saut du processus est borné par celui d'un processus de Yule. Une fois la condition initiale atteinte, nous obtenons une séquence ordonnée d'ancêtres potentiels x_1, ..., x_n qui vérifie la propriété suivante.

"Si au moins l'un des ancêtres potentiels x_1, ..., x_n au temps 0 est de type 1, alors l'individu en x au temps présent est aussi de type 1, et est un descendant du premier ancêtre de type 1 dans la séquence ordonnée. De plus, il porte le même marquage que son ancêtre au temps 0."

Formellement, le dual que nous considérons est défini sur l'espace d'états

$$\Lambda^{exp} := \left\{ \left(A^{(1)}, ..., A^{(m)} \right) : m \in \mathbb{N}\backslash\{0\} \text{ et pour tout } l \in [\![1, m]\!],\ A^{(l)} \in \mathrm{Seq}_f \left(L^{-1}\mathbb{Z} \times [\![1, M_n]\!] \right) \right\}$$

où $\mathrm{Seq}_f(L^{-1}\mathbb{Z} \times [\![1, M_n]\!])$ est l'ensemble des *séquences* finies d'éléments de $L^{-1}\mathbb{Z} \times [\![1, M_n]\!]$ dans lesquelles chaque élément n'apparaît qu'une fois.

Pour reconstruire les généalogies des individus en x_1, ..., x_m au temps t, nous partons de la condition initiale $((x_1), ..., (x_m))$. A chaque événement de reproduction, nous mettons à jour la séquence des ancêtres potentiels de chaque individu affecté, et nous remontons ainsi jusqu'à la condition initiale. A noter qu'un même individu peut apparaître simultanément dans la séquence d'ancêtres potentiels de plusieurs individus différents de l'échantillon.

Ainsi défini, ce processus vérifie les propriétés suivantes.

Proposition 1.3.1. *Soit $m \in \mathbb{N}\backslash\{0\}$, et soient $x_1, ..., x_m \in L_n^{-1}\mathbb{Z} \times [\![1, M_n]\!]$. Soit $(A_t^{(1)}, ..., A_t^{(m)})_{t\geq 0}$ le processus dual de condition initiale $((x_1), ..., (x_m))$ introduit plus haut.*

- *Pour tout $t \geq 0$,*

$$\mathbb{E}\left[\prod_{l=1}^{m} \xi_t^n(x_l) \right] = \mathbb{E}\left[\prod_{l=1}^{m} \left(1 - \prod_{x \in A_t^{(l)}} (1 - \xi_0^n(x)) \right) \right].$$

- *Pour tout* $t \geq 0$*, si nous posons* $A_t^{(l)} = (x_{t,1}^{(l)}, ..., x_{t,|A_t^{(l)}|}^{(l)})$ *pour tout* $l \in [\![1, m]\!]$*, alors*

$$\mathbb{E}\left[\eta_t^n(x_1)\right] = \mathbb{E}\left[\sum_{k=1}^{|A_t^{(1)}|} \eta_0^n(x_{t,k}^{(1)}) \prod_{k'=1}^{k-1} \left(1 - \xi_0^n(x_{t,k'}^{(1)})\right)\right]$$

et

$$\mathbb{E}\left[\prod_{l=1}^{m} \eta_t^n(x_l)\right] = \mathbb{E}\left[\prod_{l=1}^{m} \left(\sum_{k=1}^{|A_t^{(l)}|} \eta_0^n(x_{t,k}^{(l)}) \prod_{k'=1}^{k-1} \left(1 - \xi_0^n(x_{t,k'}^{(l)})\right)\right)\right].$$

Ces relations peuvent être interprétées de la façon suivante.

- L'échantillon ne contient que des individus réels si, et seulement si chacun des individus admet au moins un ancêtre potentiel réel.
- L'échantillon ne contient que des individus de type 1 marqués (i.e, de type $1*$) si et seulement si le vrai parent de chaque individu est de type 1 et marqué. Ici, la somme est faite sur tous les indices possibles pour le vrai parent.

Remarque 1.3.2. Utiliser un "type 0" pour représenter des zones vides est une idée qui apparaît dans d'autres modèles de systèmes de particules en interaction, comme par exemple le processus de contact. Le fait de les désigner sous le nom d'"individus fantômes" permet d'avoir une image plus visuelle du modèle, mais il ne faut pas perdre de vue que ces individus sont avant tout un artefact de modélisation. En particulier, la diversité génétique effectivement observable est la diversité *parmi les individus de type* 1, ou individus réels. Ceci a des conséquences en termes de distribution des généalogies : ainsi, l'ensemble des ancêtres potentiels d'un individu de type 1 est conditionné à contenir au moins un individu réel, et a une distribution différente de celle de l'ensemble des ancêtres potentiels d'un individu de type inconnu (et donc potentiellement fantôme).

Cette approche a permis d'obtenir de premiers résultats sur la dynamique de l'expansion et sur l'évolution de la diversité génétique au front. Afin de les présenter, introduisons la densité en individus de type 1 dans le patch $i \in L_n^{-1}\mathbb{Z}$ au temps t, notée $u_t^n(i)$. Nous posons ainsi

$$u_t^n(i) := \frac{1}{M_n} \sum_{j=1}^{M_n} \xi_t^n(i,j)$$

Nous prolongeons $u_t^n(\cdot)$ à $\mathbb{R}$ en interpolant entre les points de $L_n^{-1}\mathbb{Z}$.

Soit $\mathcal{C}_{[0,1]}(\mathbb{R})$ l'ensemble des fonctions $f : \mathbb{R} \to [0,1]$ continues, muni de la topologie de la convergence uniforme sur les compacts. Le résultat suivant correspond au Théorème 1 de [DF16].

Theorème 1. *Supposons que* u_0^n *converge dans* $\mathcal{C}_{[0,1]}(\mathbb{R})$ *vers* $f_0 \in \mathcal{C}_{[0,1]}(\mathbb{R})$*. Supposons de plus que*

$$\begin{aligned} r_n M_n L_n^{-2} &\xrightarrow[n\to+\infty]{} \alpha \in (0,\infty), \\ M_n R_n^{-1} &\xrightarrow[n\to+\infty]{} \beta \in [0,\infty), \\ r_n L_n^{-1} &\xrightarrow[n\to+\infty]{} \gamma \in [0,\infty), \\ L_n &\xrightarrow[n\to+\infty]{} +\infty, \\ \textit{et } L_n R_n &\xrightarrow[n\to+\infty]{} +\infty. \end{aligned}$$

Alors, la densité en individus de type 1 $(u_t^n)_{t\geq 0}$ *converge en distribution dans* $D([0,\infty), \mathcal{C}_{[0,1]}(\mathbb{R}))$ *vers un processus continu* $(u_t)_{t\geq 0}$ *à valeurs dans* $\mathcal{C}_{[0,1]}(\mathbb{R})$ *qui est la solution faible de l'équation différentielle stochastique*

$$\begin{cases} \partial_t u &= \alpha\Delta u + 2\theta\beta u(1-u) + \sqrt{4\gamma u(1-u)}\dot{W}, \\ u_0 &= f_0, \end{cases}$$

où $\dot{W}$ *est un bruit blanc espace-temps sur* $[0,\infty)\times\mathbb{R}$.

Ainsi, dans la limite en grande population, sous un rescaling diffusif et en supposant que l'avantage sélectif des individus réels sur les individus fantômes est faible (i.e, que les individus réels envahissent lentement l'espace), nous retrouvons l'équation de Fisher-KPP. Ceci est d'ailleurs un indicateur du fait que l'approche ne sera pas directement applicable au cas de populations vivant en dimension ≥ 2, et motive la construction et l'étude d'autres modèles de populations en expansion.

Le Théorème 4 de [DF16] décrit de plus les dynamiques couplées des densités d'individus de type 1 et de type 1 marqué dans la population en expansion. La preuve de ces deux théorèmes repose sur une adaptation d'une technique de preuve issue de [MS95], initialement appliquée à un modèle du votant avec interactions à longue portée.

Le résultat obtenu dans [DF16] a depuis été étendu au cas de graphes dans [Fan21].

1.3.2 Le ∞-parent SLFV

Intégration des individus fantômes au SLFV

L'intégration d'individus fantômes au SLFV peut se faire en considérant que les zones occupées contiennent des individus de type 1 (correspondant au type A dans la Section 1.2), tandis que les zones vides contiennent des individus "fantômes" de type 0 (correspondant cette fois-ci aux individus de type a dans la Section 1.2) présentant un désavantage sélectif face aux individus de type 1. Ainsi, informellement, la densité $\omega_t : \mathbb{R}^d \to [0,1]$ représente cette fois-ci les zones *vides* au temps t, modélisées comme contenant des individus fantômes. Nous obtenons alors un cas particulier du SLFV avec sélection introduit dans [FP17], auquel je me référerai sous le nom de k*-parent SLFV*.

Définition informelle 1.3.3. *(k-parent SLFV) Soient* $k\in\mathbb{N}\backslash\{0,1\}$, μ *une mesure* σ*-finie sur* $(0,+\infty)$ *vérifiant la condition (1.2.1), et* Π *un processus ponctuel de Poisson défini sur* $\mathbb{R}_+\times\mathbb{R}^d\times(0,+\infty)$ *d'intensité* $dt\otimes dx\otimes\mu(dr)$. *Soit de plus* $\omega^0 : \mathbb{R}^d\to[0,1]$ *une fonction mesurable. Le processus* Λ*-Fleming Viot spatial à* k *parents* $(\omega_{k,t})_{t\geq 0}$ *de condition initiale* ω^0 *(ou* k*-parent SLFV) est défini informellement de la façon suivante. Pour tout* $(t,x,\mathcal{R})\in\Pi$, *sachant* $\omega_{k,t-}$, k *parents potentiels sont choisis uniformément au hasard dans la boule* $\mathcal{B}(x,\mathcal{R})$ *:*

- *Si au moins l'un des* k *parents potentiels est de type* 1, *alors pour tout* $y\in\mathcal{B}(x,\mathcal{R})$, *nous posons* $\omega_{k,t}(y)=0$. *Nous considérerons de plus que le vrai parent de l'événement de reproduction est le premier parent potentiel de type* 1 *choisi. Ceci se produit avec probabilité*

$$1-\mathrm{Vol}(\mathcal{B}(x,\mathcal{R}))^{-k}\left(\int_{\mathcal{B}(x,\mathcal{R})}\omega_{k,t-}(y)dy\right)^k.$$

- *Sinon, pour tout* $y\in\mathcal{B}(x,\mathcal{R})$, $\omega_{k,t}(y)=1$, *et le vrai parent est le dernier parent potentiel choisi (qui est alors de type* 0*). Ceci se produit avec probabilité*

$$\mathrm{Vol}(\mathcal{B}(x,\mathcal{R}))^{-k}\left(\int_{\mathcal{B}(x,\mathcal{R})}\omega_{k,t-}(y)dy\right)^k.$$

Voir la Figure 1.11 pour une illustration de la dynamique.

Pour la même raison que précédemment, cette définition du k-parent SLFV n'est pas rigoureuse, car un nombre infini d'événements de reproduction se produit à chaque pas de temps dans $\mathbb{R}^d$. Formellement, le k-parent SLFV est là encore un processus à valeurs mesures, défini sur l'espace $\widetilde{\mathcal{M}}_\lambda$ introduit plus haut. Observons de plus que de la même façon que précédemment, si $\omega^0 \in \{0,1\}$, alors pour tout $t \geq 0$, $\omega_{k,t} \in \{0,1\}$ presque sûrement. Ainsi, en prenant une condition initiale adaptée, il est possible de considérer que le k-parent SLFV est défini sur l'espace $\mathcal{M}_\lambda \subset \widetilde{\mathcal{M}}_\lambda$ des mesures $M \in \widetilde{\mathcal{M}}_\lambda$ telles qu'il existe $\omega : \mathbb{R}^d \to \{0,1\}$ (et non pas $[0,1]$) mesurable qui satisfait

$$M(dx, d\kappa) = ((\omega(x)\delta_0(d\kappa) + (1-\omega(x))\delta_1(d\kappa))dx.$$

C'est ce qui est fait dans le Chapitre 3. Ceci permet entre autres de construire le k-parent SLFV à partir du processus dual encodant les généalogies, à la manière de [VW15].

Comme vu plus haut dans le cas d'autres modèles avec sélection, le processus dual est différent de celui associé au SLFV sans sélection. En effet, comme la détermination des vrais parents des événements de reproduction dépend des types des k parents potentiels, et que ces types sont inconnus, il est nécessaire de reconstruire les généalogies de chacun des k parents potentiels. Ceci permet d'en déduire leur type, et par conséquent lequel d'entre eux est le vrai parent.

Le processus dual associé au k-parent SLFV, auquel je ferai référence dans la suite sous le nom de *processus ancestral à k parents*, est défini de la façon suivante.

Définition 1.3.4. *(Processus ancestral à k parents) Soit $\Xi^0 \in \mathcal{M}_p(\mathbb{R}^d)$. Le processus ancestral à k parents $(\Xi_{k,t})_{t\geq 0}$ de condition initiale Ξ^0 est le processus markovien de saut à valeurs dans $\mathcal{M}_p(\mathbb{R}^d)$ défini de la façon suivante. Posons d'abord $\Xi_{k,0} = \Xi^0$. Puis, soit $\overleftarrow{\Pi}$ un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0,+\infty)$ d'intensité $dt \otimes dx \otimes \mu(d\mathcal{R})$. Pour tout $(t,x,\mathcal{R}) \in \overleftarrow{\Pi}$ affectant au moins un atome de $\Xi_{k,t-}$, et sachant $\Xi_{k,t-} = \sum_{i=1}^{N_{t-}} \delta_{x_i}$,*

- *Soient $y_1, ..., y_k \in \mathcal{B}(x,\mathcal{R})$ choisis indépendamment et uniformément au hasard dans $\mathcal{B}(x,\mathcal{R})$.*
- *Rappelons que $\mathcal{S}^{\mathcal{R}}(\Xi_{k,t-})$ désigne l'ensemble des entiers $i \in [\![1, N_{t-}]\!]$ tels que $x_i \in \mathcal{B}(x,\mathcal{R})$.*
- *Posons alors*

$$\Xi_{k,t} = \Xi_{k,t-} - \sum_{i \in \mathcal{S}^{\mathcal{R}}(\Xi_{k,t-})} \delta_{x_i} + \sum_{i=1}^{k} \delta_{y_i}.$$

Ce processus est bien défini, car son taux de saut peut être borné par celui d'un processus de Yule à k descendants.

Nous avons alors la relation de dualité suivante.

Proposition 1.3.5. *Soient $l \in \mathbb{N}^*$ et ψ une densité sur $(\mathbb{R}^d)^l$. Si*

$$(\Xi_{k,t})_{t\geq 0} = \left(\sum_{i=1}^{N_{k,t}} \delta_{\xi^i_{k,t}} \right)_{t\geq 0}$$

est un processus ancestral à k parents, alors pour tout $t \geq 0$,

$$\mathbb{E}_{\omega_{k,0}=\omega^0} \left[\int_{(\mathbb{R}^d)^l} \psi(x_1,...,x_l) \left\{ \prod_{j=1}^{l} \omega_{k,t}(x_j) \right\} dx_1...dx_l \right]$$
$$= \int_{(\mathbb{R}^d)^l} \psi(x_1,...,x_l) \mathbb{E}_{\Xi_{k,0}=\Xi[x_1,...,x_l]} \left[\prod_{j=1}^{N_{k,t}} \omega^0(\xi^j_{k,t}) \right] dx_1...dx_l.$$

Cette relation peut être interprétée de la façon suivante. La probabilité que l individus situés en x_1, ..., x_l au temps t soient de type 0 (i.e, que les positions spatiales correspondantes soient vides) est égale à la probabilité que l'*ensemble* des ancêtres potentiels de chacun d'entre eux soient de type 0. Ceci justifie d'ailleurs a posteriori de suivre les densités en individus de type 0 plutôt qu'en individus de type 1. En effet, la formulation de la relation de la dualité est alors plus simple et lisible. La preuve de la relation de dualité ainsi que de la construction du k-parent SLFV sont des adaptations directes des preuves dans [EVY20], qui traite du cas $k = 2$.

Limite du k-parent SLFV lorsque $k \to +\infty$

L'interprétation du type 0 comme correspondant à des individus fantômes et à des zones vides nous encourage à explorer la limite du k-parent SLFV lorsque $k \to +\infty$. Ce régime de paramètres correspond à un très fort avantage sélectif des individus réels face aux individus fantômes. La dynamique du processus limite est la suivante. A chaque événement de reproduction :

- Si la zone affectée contient une fraction non nulle d'individus de type 1, alors le vrai parent est choisi uniformément au hasard parmi tous les individus de type 1, et ses descendants remplissent entièrement la zone.
- Sinon, la zone reste entièrement vide.

D'un point de vue modélisation, l'intérêt de ce processus vient du fait qu'il ressemble à un modèle d'Eden, mais continu en espace, et associé à un processus dual encodant les généalogies. Ceci suggère donc de s'intéresser aux propriétés de croissance du ∞-parent SLFV, afin de les comparer au modèle d'Eden, comme nous le ferons dans la prochaine section et dans le Chapitre 4.

D'un point de vue mathématique, la particularité principale de ce processus par rapport aux autres SLFVs est directement liée à la forme de son processus dual. En effet, à chaque événement de reproduction, il faut maintenant connaître l'intégralité de la composition de la zone affectée afin d'en déduire le vrai parent. Ainsi, au lieu d'être défini sur l'espace $\mathcal{M}_p(\mathbb{R}^d)$ des mesures ponctuelles, le dual du ∞-parent SLFV est cette fois-ci défini sur ce qui peut être interprété comme l'espace des formes géométriques sur $\mathbb{R}^d$: l'ensemble $\mathcal{M}^{cf}$ des mesures de la forme

$$\mathcal{M}^{cf} := \{m(E) = \mathbb{1}_{\{x \in E\}} dx : E \in \mathcal{E}^{cf}\},$$

où $\mathcal{E}^{cf}$ est l'ensemble des unions finies des sous-ensembles Lebesgue-mesurables, fermés, connexes et de mesure de Lebesgue finie et non nulle de $\mathbb{R}^d$. Dans le cas du dual du ∞-parent SLFV, ces sous-ensembles seront d'ailleurs généralement des boules.

Le dual du ∞-parent SLFV, appelé le *processus ancestral à ∞ parents*, est défini de la façon suivante, étant donné une mesure σ-finie μ sur $(0, \infty)$ satisfaisant la condition (1.2.1).

Définition 1.3.6. *(Processus ancestral à ∞ parents) Soit $E^0 \in \mathcal{E}^{cf}$. Le processus ancestral à ∞ parents $(m(E_t))_{t \geq 0}$ de condition initiale $m(E^0)$ est le processus markovien de saut à valeurs dans $\mathcal{M}^{cf}$ défini de la façon suivante. Posons d'abord $m(E_0) = m(E^0)$. Puis, soit $\overleftarrow{\Pi}$ un processus ponctuel de Poisson sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0, +\infty)$ d'intensité $dt \otimes dx \otimes \mu(d\mathcal{R})$. Pour tout $(t, x, \mathcal{R}) \in \overleftarrow{\Pi}$, si $\mathrm{Vol}(\mathcal{B}(x, \mathcal{R}) \cap E_{t-}) \neq 0$, alors*

$$m(E_t) = m(E_{t-} \cup \mathcal{B}(x, \mathcal{R})).$$

Afin de vérifier si ce processus est bien défini ou non, il n'est maintenant plus possible de comparer son taux de saut à celui d'un processus de Yule de la façon dont nous l'avions fait précédemment. En effet, lorsque k est fini, nous pouvions utiliser le fait que chaque atome était affecté par un événement de reproduction à un taux identique. Ici, l'analogue des atomes serait les boules

associées à chaque événement de reproduction, mais le taux auquel elles sont affectées par un événement de reproduction dépend de leur rayon.

Dans le cas où le rayon des boules de reproduction est borné par $\mathcal{R}_0 > 0$ (i.e, si $\mu((\mathcal{R}_0, \infty)) = 0$), il est quand même possible d'adapter la preuve en considérant le taux maximal auquel une boule est affectée par un événement de reproduction plutôt que le taux auquel un atome est affecté. Ceci permet de borner le taux de saut du processus par celui d'un processus de Yule à deux descendants, et donc de conclure que le processus ancestral à ∞ parents est bien défini.

En revanche, si le rayon des boules de reproduction n'est pas borné, alors le processus n'est pas forcément bien défini. Si μ satisfait une certaine hypothèse technique (introduite dans le Chapitre 3), il est possible de comparer le nombre de sauts du processus sur l'intervalle $[0, t)$ au nombre de particules dans un certain processus de branchement, dont la condition initiale dépend à la fois de E^0 et de μ. La non-explosion du processus de branchement implique alors la non-explosion du processus ancestral à ∞ parents. Mais le contraire n'est pas forcément vrai et ainsi, rien ne permet de conclure au caractère bien défini du processus lorsque μ ne satisfait pas l'hypothèse technique.

Afin de construire rigoureusement le ∞-parent SLFV, il est possible de s'appuyer sur la structure de preuve introduite dans la Section 1.2.3 :

1. Définir un opérateur correspondant à la dynamique du processus, et construire une solution au problème martingale associé à cet opérateur en prenant la limite d'une séquence de ∞-parent SLFV, définis sur une suite de compacts de $\mathbb{R}^d$ via le processus ponctuel de Poisson encodant les événements de reproduction.

2. Construire le processus dual candidat.

3. Établir une relation de dualité entre toute solution au problème martingale et le processus dual candidat, afin de conclure à l'unicité de la solution au problème martingale.

La relation de dualité liant le ∞-parent SLFV à son dual a la forme suivante, assez différente de celle correspondant aux autres SLFVs.

Proposition 1.3.7. *Soient $\omega^0 : \mathbb{R}^d \to \{0, 1\}$ mesurable et $E^0 \in \mathcal{E}^{cf}$. Soient $(\omega_{\infty,t})_{t\geq 0}$ le ∞-parent SLFV de condition initiale ω^0 et $(m(E_t))_{t\geq 0}$ le processus ancestral à ∞ parents de condition initiale E^0. Alors, pour tout $t \geq 0$,*

$$\mathbb{E}\left[\delta_0\left(\int_{E^0}(1-\omega_{\infty,t}(x))\,dx\right)\right] = \mathbb{E}\left[\delta_0\left(\int_{E_t}\left(1-\omega^0(x)\right)dx\right)\right],$$

où $\delta_0 : x \in \mathbb{R} \to \{0, 1\}$ est la fonction égale à 1 si $x = 0$, et 0 sinon.

En d'autres termes, la zone E^0 est entièrement vide au temps t si, et seulement si les ancêtres potentiels au temps 0 des individus occupant la zone au temps t sont tous de type 0.

A noter que du fait de l'étape 2 de la preuve, la construction du ∞-parent SLFV que je viens de présenter n'est valable que lorsque μ satisfait l'hypothèse technique supplémentaire mentionnée plus haut. Contrairement aux autres SLFVs, le ∞-parent SLFV ne peut donc être défini comme l'unique solution d'un problème martingale que sous une condition plus restrictive sur μ. Il est cependant quand même possible de définir autrement le ∞-parent SLFV, à partir d'une suite de k-parent SLFVs couplés (voir le Chapitre 3, Section 2). Cette construction alternative a deux intérêts :

- elle reste valable même lorsque μ ne vérifie pas l'hypothèse technique,
- elle permet de faire apparaître de façon plus explicite le ∞-parent SLFV comme limite du k-parent SLFV lorsque $k \to +\infty$.

Le couplage entre k-parent SLFVs repose sur la même idée de processus ponctuel de Poisson étendu introduite plus haut et issue de [VW15]. Précédemment, nous avons vu que dans le cas du SLFV sans sélection, il était possible d'enrichir les informations données sur les événements de reproduction par le processus ponctuel de Poisson, en ajoutant la localisation du parent de chaque événement de reproduction. Il est possible de généraliser cette approche au k-parent SLFV en intégrant les positions des k parents potentiels choisis. Pour coupler différents k-parents SLFVs entre eux via l'utilisation d'un même processus ponctuel de Poisson étendu, l'approche que j'ai utilisée est la suivante : associer à chaque événement de reproduction une suite $(\mathcal{P}_n)_{n\geq 1} \in \mathcal{B}(0,1)^{\mathbb{N}}$ de positions dans la boule de reproduction (correspondant à des parents potentiels), chacune distribuée uniformément au hasard dans $\mathcal{B}(0,1)$ et indépendamment des autres positions. Notons ainsi $\tilde{u}$ la loi suivie par $(\mathcal{P}_n)_{n\geq 1}$. Le processus ponctuel de Poisson que nous considérons est alors défini sur $\mathbb{R}_+ \times \mathbb{R}^d \times (0,\infty) \times \mathcal{B}(0,1)^{\mathbb{N}}$ et d'intensité $dt \otimes dx \otimes \mu(d\mathcal{R}) \otimes \tilde{u}(d(p_n))_{n\geq 1}$. Ce processus ponctuel de Poisson peut ensuite être utilisé pour construire une suite de k-parent SLFVs, $k \geq 2$ tous de même condition initiale, en utilisant à chaque fois les k *premiers parents potentiels* tirés pour obtenir le k-parent SLFV. Le couplage fait en sorte que toute zone occupée dans le k-parent SLFV au temps t l'est aussi dans le k'-parent SLFV, $k' \geq k$. Formellement, cela revient à dire que

$$\forall t \geq 0, \forall x \in \mathbb{R}^d, \forall k, k' \in \mathbb{N}, k' \geq k \Longrightarrow \omega_{k',t}(x) \leq \omega_{k,t}(x).$$

Ainsi, pour tout $t \geq 0$ et $x \in \mathbb{R}^d$, la suite $(\omega_{k,t}(x))_{k\geq 2}$ est décroissante. Etant à valeurs dans $\{0,1\}$, elle converge vers $\omega_{\infty,t}(x) \in \{0,1\}$. Le ∞-parent SLFV est alors défini de la façon suivante.

Définition (semi) informelle 1.3.8. *(∞-parent SLFV) Le ∞-parent SLFV $(M_t)_{t\geq 0}$ est l'unique processus markovien à valeurs dans $\mathcal{M}_\lambda$ tel que pour tout $t \geq 0$, M_t satisfait*

$$M(dx, d\kappa) = ((\omega_{\infty,t}(x)\delta_0(d\kappa) + (1 - \omega_{\infty,t}(x))\delta_1(d\kappa))dx.$$

Vitesse de croissance et lien avec le modèle d'Eden

Dans cette section, nous nous focalisons sur le cas de la dimension 2.

D'un point de vue modélisation, l'intérêt du ∞-parent SLFV provient du fait qu'il semble être un équivalent continu en espace du modèle d'Eden, qui permet de plus l'étude de la diversité génétique. Pour autant, il s'agit avant tout d'une conjecture, et les propriétés des expansions générées par un ∞-parent SLFV sont peut-être en fait très différentes de celles des expansions générées par un modèle d'Eden.

L'objet du Chapitre 4 est ainsi l'étude des propriétés de croissance de la région occupée par les individus réels dans le ∞-parent SLFV, afin de les comparer à celles du modèle d'Eden. Pour cela, rappelons d'abord quels sont les résultats connus sur le modèle d'Eden, et plus généralement sur les modèles de percolation de premier passage. D'après le théorème de forme de Cox-Durrett [CD81], l'expansion d'une population suivant le modèle d'Eden est linéaire en temps. Il est possible d'obtenir des bornes supérieures et inférieures sur la vitesse de croissance (voir par exemple [AP02; BK93]), qui sont d'ailleurs généralement aussi valables pour les modèles de percolation "à courte portée" (*short-range percolation*) lorsque le temps de passage d'une arête du graphe sous-jacent est presque sûrement non nul (voir par exemple [ADH17]). Il n'existe cependant pas de formule explicite pour la vitesse d'expansion du modèle d'Eden, et il est nécessaire d'utiliser des simulations pour en obtenir une approximation [AD15]. Ceci n'est d'ailleurs pas spécifique au modèle d'Eden, mais est en fait aussi le cas pour la quasi-intégralité des modèles issus de la théorie de la percolation, le *corner growth model* (voir [Sep09]) étant en quelque sorte l'exception qui confirme la règle.

Le modèle d'Eden est de plus conjecturé comme appartenant à la classe d'équivalence de l'équation KPZ, pour laquelle les fluctuations spatiales de la position du front parallèlement au demi-plan duquel part l'expansion augmentent initialement en $t^{1/3}$. Là encore, il n'existe pas de preuve

rigoureuse de ce résultat, ni pour le modèle d'Eden, ni pour la quasi-intégralité des modèles probabilistes de croissance de population (à l'exception notable du *Solid on Solid growth model*, voir [BG97]).

Ces rappels sur le modèle d'Eden ont deux intérêts : illustrer quelles propriétés fondamentales doivent être vérifiées par le ∞-parent SLFV pour que nous puissions raisonnablement le considérer comme un équivalent du modèle d'Eden, et donner une idée du type de résultats que nous pouvons raisonnablement espérer démontrer. Dans le Chapitre 4, je montre ainsi que l'expansion de l'aire occupée par les individus de type 1 dans le ∞-parent SLFV est linéaire en temps, de la même façon que pour le modèle d'Eden. Plus précisément, je prouve le résultat suivant lorsqu'il existe $\mathcal{R}_0 > 0$ tel que $\mu((\mathcal{R}_0, +\infty)) = 0$, ou en d'autres termes, lorsque le rayon des événements de reproduction est borné.

Theorème 2. *Supposons que la population d'individus réels occupe initialement le demi-plan*

$$\overline{HP}^0 := \{(x, y) \in \mathbb{R}^2 : x < 0\},$$

et considérons le ∞-parent SLFV $(M_t^{HP})_{t\geq 0}$ de condition initiale

$$M_0^{HP}(dz) := \mathbb{1}_{\{z\in\overline{HP}^0\}}dz.$$

Soit $(\omega_t^{HP})_{t\geq 0}$ une densité de $(M_t^{HP})_{t\geq 0}$, et pour tout $x \geq 0$, posons

$$\overrightarrow{\tau}_x := \min\left\{t \geq 0 : \lim_{\epsilon\to 0} V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz < 1\right\},$$

où $\mathcal{B}_\epsilon((x, 0))$ est la boule de centre $(x, 0)$ et de rayon ϵ, et où V_ϵ est le volume de $\mathcal{B}_\epsilon((x, 0))$. Alors, il existe $\nu > 0$ tel que

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overrightarrow{\tau}_x]}{x} = \nu.$$

Intuitivement, $\overrightarrow{\tau}_x$ peut être interprété comme le temps auquel le point $(x, 0)$ est atteint par le processus. Ce théorème dit donc que lorsque les événements de reproduction ont un rayon borné, l'avancée des individus réels le long d'une ligne est linéaire en temps. La vitesse correspondante est égale à ν^{-1}, qui n'a pas d'expression explicite. Ce résultat reste vrai si l'on remplace les boules de reproduction par des *ellipses*, et pourrait facilement être généralisé à d'autres formes géométriques.

La preuve de ce résultat repose sur l'utilisation du processus dual associé au ∞-parent SLFV, le processus ancestral à ∞ parents. En effet, une position de l'espace est occupée par des individus de type 1 si, et seulement si au moins l'un de ses ancêtres potentiels au temps 0 appartient au demi-plan $\overline{HP}^0$. En utilisant l'invariance par translation et rotation du processus ponctuel de Poisson encodant les événements de reproduction et associé au processus dual, le théorème précédent est alors équivalent au résultat suivant.

Proposition 1.3.9. *Pour tout $x > 0$, soit HP^x le demi-plan :*

$$HP^x := \{(x', y) \in \mathbb{R}^2 : x' \geq x\}.$$

Soit de plus $(E_t)_{t\geq 0}$ le processus ancestral à ∞ parents de condition initiale $\{(0, 0)\}$ (en modifiant la définition du processus de façon à pouvoir avoir un point comme condition initiale), et posons

$$\overleftarrow{\tau}_x := \min\{t \geq 0 : \mathrm{Vol}(E_t \cap HP^x) > 0\}.$$

Alors, il existe $\nu > 0$ tel que

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} = \nu.$$

La preuve de ce résultat comporte deux parties. Dans un premier temps, nous montrons que l'expansion est *au moins* linéaire en temps, i.e qu'il existe $\nu \geq 0$ (et non pas $\nu > 0$) tel que

$$\lim_{x \to +\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} = \nu.$$

Pour cela, nous nous intéressons aux temps d'atteinte des demi-plans $(HP^{4\mathcal{R}_0 n})_{n\in\mathbb{N}}$, et pour tout $n \in \mathbb{N}$, nous posons

$$T_{0,n} = \overleftarrow{\tau}_{4n\mathcal{R}_0}.$$

Ce choix est justifié par le fait que comme les événements de reproduction ont un rayon borné par $\mathcal{R}_0$, il n'est (presque sûrement) pas possible d'atteindre simultanément les demi-plans $HP^{4n\mathcal{R}_0}$ et $HP^{4n'\mathcal{R}_0}$, $n \neq n'$.

Puis, pour tout $m \in \mathbb{N}$, à l'instant où le demi-plan $HP^{4m\mathcal{R}_0}$ est atteint, nous choisissons un point de la forme $(4m\mathcal{R}_0, y)$, $y \in \mathbb{R}$ dans l'ensemble $E_{T_{0,m}} \cap HP^{4m\mathcal{R}_0}$, et redémarrons un processus ancestral à ∞ parents de ce point. Le processus ancestral est construit en prenant le même processus ponctuel de Poisson sous-jacent et en ne considérant que les événements de reproduction qui se produisent après l'instant $T_{0,m}$. Ceci est possible par invariance par translation dans le temps de la distribution du processus ponctuel de Poisson.

Notons alors $T_{m,n}$, $n \geq m$ le temps nécessaire à ce processus ancestral pour atteindre le demi-plan $HP^{4n\mathcal{R}_0}$. Ces temps vérifient

$$\forall n \in \mathbb{N}, n \geq m \Longrightarrow T_{0,n} \leq T_{0,m} + T_{m,n}.$$

La famille $(T_{m,n})_{0\leq m\leq n}$ vérifie de plus les autres hypothèses du Théorème 1.10 de [Lig85], ce qui permet d'obtenir l'existence de $\nu' \geq 0$ tel que

$$\lim_{n \to +\infty} \frac{\mathbb{E}[T_{0,n}]}{n} = \nu'.$$

Nous concluons alors en utilisant le fait que $(\overleftarrow{\tau}_x)_{x\geq 0}$ est croissante. Montrer que $(T_{m,n})_{0\leq m\leq n}$ vérifie les hypothèses du Théorème 1.10 donne de plus une borne inférieure sur la vitesse de croissance (qui correspond à une borne supérieure sur ν).

La deuxième partie de la preuve consiste cette fois-ci à montrer que l'expansion est *au plus* linéaire en temps, en comparant le processus dual à un problème de percolation de premier passage. La comparaison repose sur un découpage de l'espace en *cellules* carrées de côté $2\mathcal{R}_0$, placées sur une grille et distantes de $6\mathcal{R}_0$. Nous utilisons ces cellules pour discrétiser le processus ancestral à ∞ parents, en considérant qu'une cellule est *active* si elle intersecte le processus ancestral, et est *inactive* sinon. Le processus ancestral partant d'un point peut être vu comme une union de boules de rayon au plus $\mathcal{R}_0$, et toutes n'intersectent pas une cellule, mais elles sont toutes forcément à une distance finie d'une cellule active. Ainsi, la vitesse de croissance du processus ancestral à ∞ parents est la même que celle du processus discrétisé de cellules actives.

C'est ce processus discrétisé que nous utilisons pour construire la comparaison avec le problème de percolation de premier passage. En effet, observons qu'une cellule ne peut devenir active que si l'une des huit cellules voisines les plus proches l'est aussi. De plus, lorsqu'une première cellule voisine devient active, il faut attendre qu'un événement de reproduction affecte la cellule d'intérêt pour qu'elle puisse potentiellement devenir active, ce qui se produit après un temps distribué selon une loi exponentielle. Ainsi, nous pouvons comparer les cellules actives et inactives du processus discrétisé aux sites vides et occupés d'un problème de percolation de premier passage "à courte portée" dans lequel chaque arête est traversée en un temps exponentiel. Nous concluons en utilisant le fait que les expansions dans ce type de problèmes de percolation sont linéaires en temps [CD16].

Afin de compléter cette étude théorique, nous avons aussi simulé le ∞-parent SLFV, afin d'obtenir une approximation de la vitesse de croissance de la zone occupée par les individus réels, et afin d'étudier les fluctuations de la position du front le long d'une ligne parallèle à la frontière du demi-plan de départ. Les simulations ont montré que la vitesse de croissance était beaucoup plus élevée qu'initialement conjecturé, du fait de la formation de "pics" au niveau du front, pointant dans la direction d'expansion. Ces pics sont relativement rares, mais lorsqu'ils se produisent, ils "épaississent" ensuite dans la direction transverse à l'expansion, ce qui conduit à une avancée soudaine du front sur de grandes distances. Voir la Figure 1.12 pour une illustration de cette dynamique. La dernière section du Chapitre 4 s'intéresse à un modèle jouet d'expansion pour comprendre comment les pics peuvent faire avancer le front plus vite. L'intérêt de ce modèle jouet est qu'il permet d'obtenir une expression explicite pour la vitesse d'expansion, en utilisant la distribution invariante d'une version discrétisée en temps du processus.

L'étude des fluctuations du front suggère qu'elles augmentent initialement en $t^{1/3}$, mais n'est pour autant pas très concluante. Une analyse plus fine de la croissance des fluctuations se heurte à des problèmes computationnels, liés à la façon dont le processus est simulé.

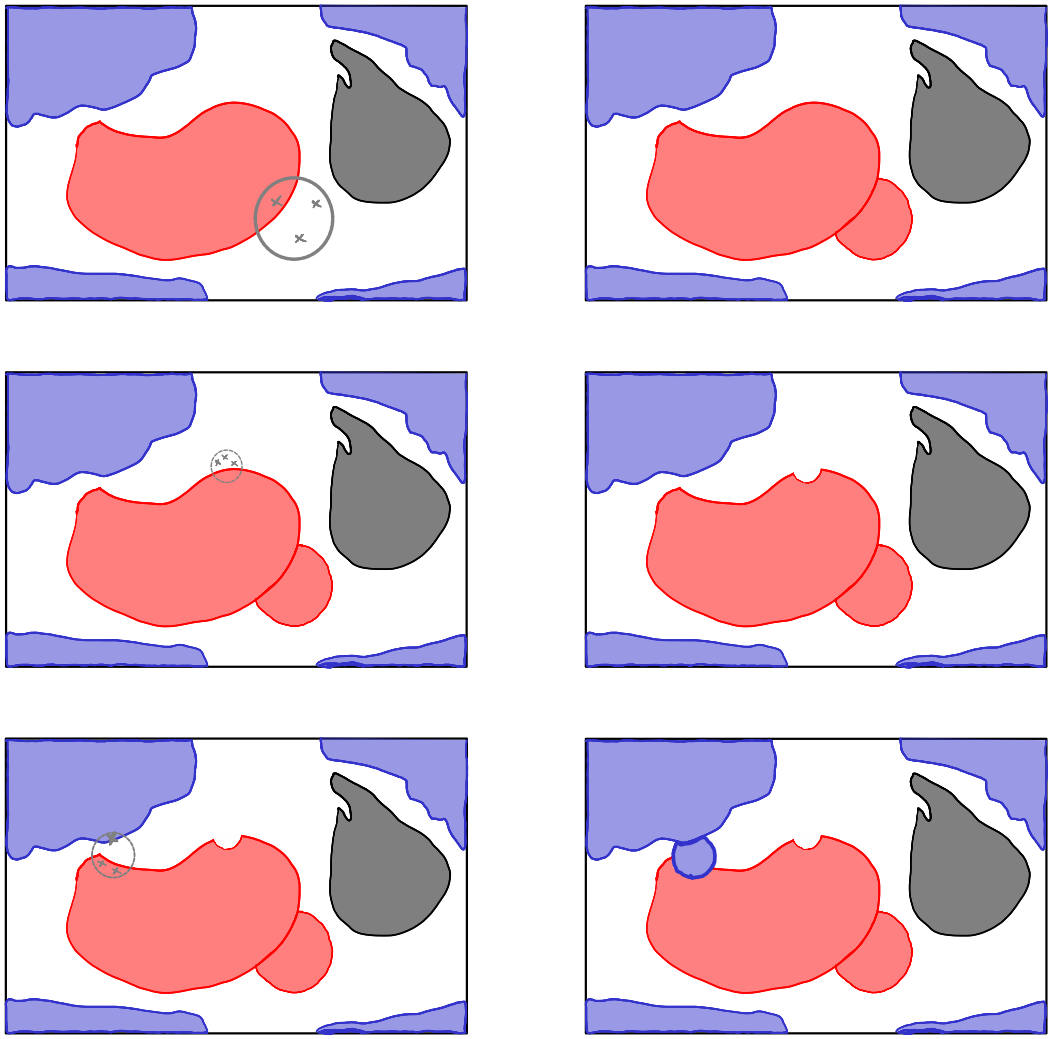


Figure 1.11: Définition heuristique du k-parent SLFV, ici dans le cas $k = 3$. Nous supposons qu'initialement, les individus occupant une même position de l'espace sont du même type. Cependant, cette fois-ci, les zones blanches sont vides, et modélisées comme occupées par des individus fantômes. (a), (c), (e) Nous attendons qu'un premier événement de reproduction affecte la zone. Nous choisissons alors 3 parents potentiels dans la zone. (b) Si un seul des 3 parents potentiels est réel, alors il s'agit du vrai parent. Il se reproduit, et ses descendants remplissent la boule. (d) Si les 3 parents potentiels sont fantômes, alors la boule est remplie d'individus fantômes. Ceci correspond à un événement d'extinction locale. (f) Si au moins deux parents potentiels sont réels, alors le vrai parent est le premier parent potentiel réel choisi. Ici, nous supposerons qu'il s'agit du parent bleu.

Figure 1.12: Croissance de la zone occupée dans le ∞-parent SLFV, à partir d'une condition initiale dans laquelle les individus réels occupent initialement le demi-plan situé juste au-dessus de l'image. L'expansion de la sous-population d'individus réels a ici lieu vers le bas, et les images représentent l'état de la même population après des durées d'expansion de plus en plus longues. Chaque couleur représente un sous-type différent parmi les individus réels. Les zones blanches correspondent aux zones vides, modélisées comme occupées par des individus fantômes.

1.3.3 Une extension du modèle de Wright-Fisher pour les métapopulations de plantes vivant dans un environnement fragmenté et perturbé

Intéressons-nous maintenant à l'intégration des individus fantômes au modèle de Wright-Fisher. Pour rappel, l'objectif est de construire une extension d'un modèle de Wright-Fisher structuré spatialement qui permette d'étudier la dynamique des plantes dans les pieds d'arbre d'alignement. Par conséquent, le modèle doit intégrer les éléments suivants :

- des événements d'extinction fréquents et localisés,
- une banque de graines.

Pour cela, commençons par introduire le modèle de Wright-Fisher structuré qui va servir de base à la définition de notre processus. Nous rajouterons ensuite les différentes composantes désirées au fur et à mesure. Considérons donc que la métapopulation est constituée d'un nombre infini de patchs indexés par $i \in \mathbb{Z}$ et situés le long d'une ligne. Chaque patch peut contenir un nombre $M \in \mathbb{N}\backslash\{0\}$ de plantes. Afin de favoriser l'intégration future de la banque de graines au modèle, nous allons supposer qu'au début de chaque génération, les M plantes sont présentes sous forme de graines. Ces M graines vont alors toutes germer, et donner des plantes, qui vont elles-mêmes produire de nouvelles graines de la façon suivante : chaque graine du patch i se choisit un parent uniformément au hasard parmi les plantes du patch i avec probabilité $1-2c$, $c \in (0, 1/2)$, ou parmi celles du patch $i-1$ (resp. $i+1$) avec probabilité c. Ceci correspond bien à un modèle de Wright-Fisher avec une structuration spatiale de type *stepping-stone* (voir la Section 1.2.2).

Commençons par intégrer la notion d'individus fantômes au modèle. Les graines et plantes peuvent maintenant être de deux types : le type 1, correspondant aux individus réellement présents, et le type 0, ou "fantôme", encodant la place libre dans un patch. La dynamique de reproduction est de plus modifiée de la façon suivante. Au lieu de choisir un seul parent, chaque graine du patch i choisir $k \in \mathbb{N}\backslash\{0,1\}$ *parents potentiels*, chacun étant pris dans le patch i avec probabilité $1-2c$ et dans le patch $i-1$ (resp. $i+1$) avec probabilité c. La graine est alors de type 1 si au moins l'un de ses parents potentiels est de type 1 (dans ce cas, le vrai parent est le premier parent potentiel de type 1 choisi), et de type 0 sinon. Comme précédemment, le paramètre k permet de quantifier l'"avantage sélectif" des individus réels sur les individus fantômes, et donc la vitesse à laquelle les individus réels arrivent à coloniser un nouveau milieu.

L'ajout des individus fantômes permet de modéliser facilement les événements d'extinction comme des "mutations" de tous les individus d'un patch vers le type 0. Nous allons supposer que les événements d'extinction se produisent après la phase de germination, mais avant la phase de production des graines. Lorsqu'un patch est affecté par un événement d'extinction, toutes les plantes qu'il contient meurent et deviennent des plantes fantômes, soit des plantes de type 0. A chaque génération, chaque patch est affecté par un événement d'extinction indépendamment des autres et avec probabilité $p \in [0, 1]$.

Il ne nous reste plus qu'à intégrer la banque de graines au modèle. Pour cela, nous allons combiner des idées des modèles de [Bla+16] et [KKL01]. En effet, nous nous intéressons à des plantes, donc à des espèces dont les graines ne peuvent généralement rester dormantes sans perdre en viabilité que sur des durées somme toute assez limitées, de l'ordre de la dizaine d'années au plus (même s'il existe des exceptions). Ceci correspond plutôt aux hypothèses du modèle de [KKL01], mais le modèle introduit dans [Bla+16] est plus facile à étudier du fait de son caractère markovien. L'utilisation des individus fantômes va permettre d'introduire l'idée d'une perte de viabilité des graines passé un certain âge au modèle de [Bla+16].

Comme précédemment, nous supposons donc que chaque patch contient exactement M graines de type 0 et 1, ou en d'autres termes, au plus M graines réelles. Chacune de ces graines peut rester

dormante pendant au plus $H \in \mathbb{N}$ générations complètes sans perdre en viabilité. Si elle reste dormante plus longtemps, alors elle "mute" et devient une graine fantôme de type 0.

Puis, à chaque génération, au lieu d'avoir germination de la banque de graines complète, seules $\lfloor gM \rfloor$ graines germent, où $g \in (0,1)$. Ces graines sont choisies uniformément au hasard dans la banque de graines, indépendamment de leur type et du temps déjà passé dans la banque de graines. Après les potentiels événements d'extinction, $\lfloor gM \rfloor$ nouvelles graines sont produites et intègrent la banque de graines, tandis que les plantes meurent.

Nous avons maintenant fini de construire notre modèle de plantes dans les pieds d'arbres d'alignement, qui est de façon plus générale adapté aux métapopulations de plantes vivant dans des environnements fragmentés et perturbés. Par la suite, je ferai référence à ce modèle sous le nom de k-parent WFSB (pour "*Wright-Fisher metapopulation process with a Seed Bank component*"). Voir la Figure 1.13 pour un résumé illustré de la dynamique du modèle.

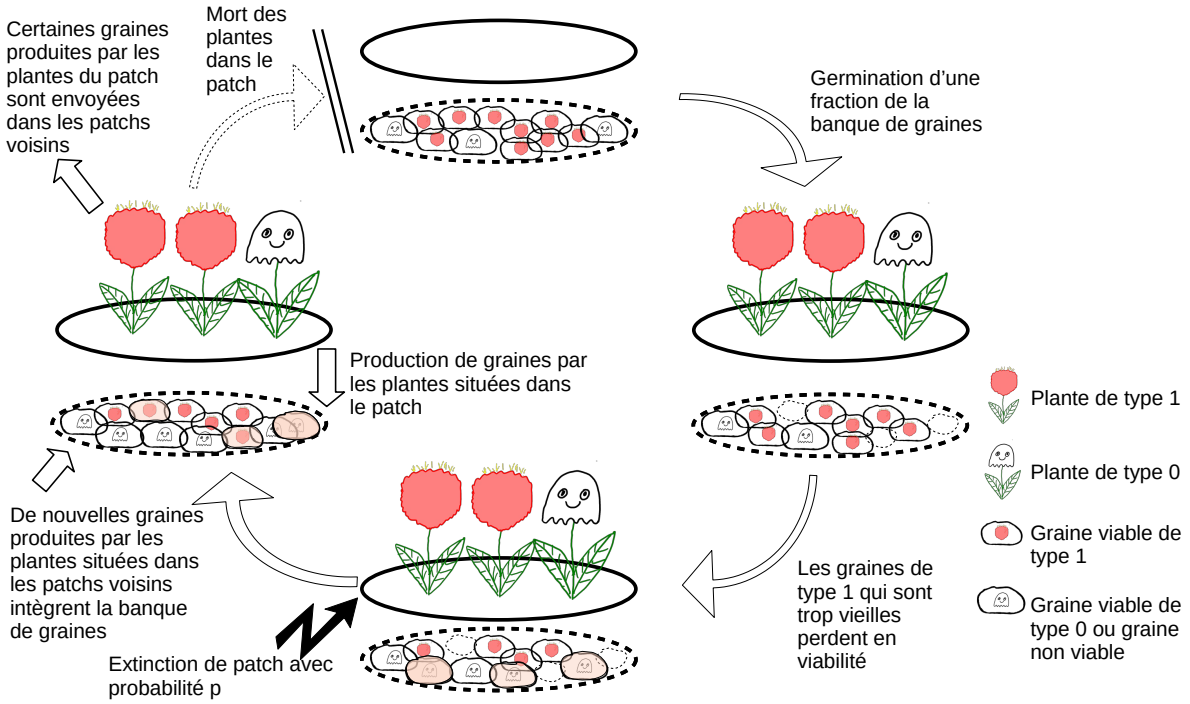


Figure 1.13: Illustration de la dynamique intra-patch du k-parent WFSB. Ici, $\lfloor gM \rfloor = 3$ et $M = 12$.

Formellement, le modèle est défini sur l'espace d'états $\mathcal{F}_M \times \mathcal{H}_M$, où

$$\mathcal{F}_M := \left\{ (\xi_{i,j})_{i\in\mathbb{Z}, j\in[\![1,M]\!]} \in \{0,1\}^{\mathbb{Z}\times[\![1,M]\!]} : \mathrm{Card}\left(\{(i,j) \in \mathbb{Z}\times[\![1,M]\!] : \xi_{i,j} = 1\}\right) < +\infty \right\},$$

and $\mathcal{H}_M := \left\{ (h_{i,j})_{i\in\mathbb{Z}, j\in[\![1,M]\!]} : \forall (i,j) \in \mathbb{Z}\times[\![1,M]\!], h_{i,j} \in \mathbb{N} \right\}$.

Pour tout $(i,j) \in \mathbb{Z}\times[\![1,M]\!]$, $h_{i,j}$ représente le nombre de générations complètes que la graine occupant le compartiment de banque de graines j du patch i a déjà passé dans la banque de graines, et $\xi_{i,j}$ représente son type *lorsqu'elle a été produite*. Autrement dit, le type actuel de la plante est $\xi_{i,j}\mathbb{1}_{\{h_{i,j}\leq H\}}$. Le modèle est alors défini de la façon suivante.

Définition 1.3.10. *Soit $(\xi,h) \in \mathcal{F}_M \times \mathcal{H}_M$. Le k-parent WFSB $(\xi^n,h^n)_{n\in\mathbb{N}}$ de condition initiale (ξ,h) est la chaîne de Markov à valeurs dans $\mathcal{F}_M \times \mathcal{H}_M$ telle que $(\xi^0,h^0) = (\xi,h)$ et pour tout $n \in \mathbb{N}$, sachant (ξ^n,h^n) :*

1. *Pour chaque $i \in \mathbb{Z}$, nous choisissons $\lfloor gM \rfloor$ compartiments de banque de graines différents $s_{i,1}, ..., s_{i,\lfloor gM \rfloor} \in [\![1, M]\!]$ uniformément au hasard dans le patch i.*

2. *Soient $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ des variables aléatoires de Bernouilli de probabilité de succès p.*

3. *Pour tout $(i, j) \in \mathbb{Z}\times[\![1, M]\!]$, si $j \in \{s_{i,j'} : j' \in [\![1, \lfloor gM \rfloor]\!]\}$, alors nous posons d'abord $h_{i,j}^{n+1} = 0$. De plus, soient $C_{1,i,j}, ..., C_{k,i,j}$ les variables aléatoires i.i.d à valeurs dans $\{-1, 0, 1\}$ telles que*

$$\mathbb{P}(C_{1,i,j} = 1) = \mathbb{P}(C_{1,i,j} = -1) = c,$$

   Pour tout $l \in [\![1, k]\!]$, si $\mathrm{Ext}_{i+C_{l,i,j}} = 1$, alors nous posons $\tilde{k}_l = 0$, et si $\mathrm{Ext}_{i+C_{l,i,j}} = 0$, nous choisissons j_l uniformément au hasard dans $\{s_{i+C_{l,i,j},j'} : j' \in [\![1, \lfloor gM \rfloor]\!]\}$, et nous posons

$$\tilde{k}_l = \xi^n_{i+C_{l,i,j},j_l} \mathbb{1}_{\{h^n_{i+C_{l,i,j},j_l} \leq H\}}.$$

   Nous posons alors $\xi_{i,j}^{n+1} = \max\{\tilde{k}_l : l \in [\![1, k]\!]\}$.

4. *Sinon, nous posons $\xi_{i,j}^{n+1} = \xi_{i,j}^n$ and $h_{i,j}^{n+1} = h_{i,j}^n + 1$.*

Il est possible de définir un processus dual associé au k-parent WFSB, en suivant là encore les ancêtres potentiels d'un individu. Cette fois-ci, il n'est cependant pas forcément nécessaire de reconstruire la généalogie complète d'un parent potentiel pour connaître son type. En effet, si la plante choisie comme parent potentiel est dans un patch qui vient d'être affecté par un événement d'extinction, alors nous savons qu'il s'agit d'une plante de type 0. De même, si la plante est issue d'une graine produite il y a plus de $H+1$ générations, alors la graine est non viable (donc de type 0) lors de sa germination, et la plante correspondante est de type 0. Le processus dual ressemble donc à l'ASG, mais auquel certaines branches sont *taillées* suite à des événements d'extinction ou à la perte de viabilité des graines. Cette construction est basée sur le *pruned ASG* introduit dans [Len+15]. Je ne définirai toutefois pas ce processus rigoureusement, les résultats du Chapitre 6 ne reposant pas sur l'utilisation du processus dual.

Dans le cadre de ma thèse, je me suis ainsi focalisée sur l'étude du modèle *forwards-in-time*, afin de comprendre sous quelles conditions une expansion est possible ou non. Pour cela, je me suis appuyée sur l'observation suivante. Si des plantes réelles sont présentes dans le patch i à la génération n, alors ces plantes peuvent produire des graines qui vont potentiellement intégrer les banques de graines des patchs $i-1$, i et $i+1$, et germer durant les générations $n+1$, ..., $n+H+1$. Nous pouvons donc définir une notion de *patchs atteignables*. En plus de cela, certains de ces patchs vont en fait être affectés par des événements d'extinction, et ne pourront pas contenir de plantes réelles. En d'autres termes, si nous nous plaçons dans l'espace $\mathbb{Z} \times \mathbb{N}$, où $(i, n) \in \mathbb{Z} \times \mathbb{N}$ correspond au patch i durant la génération n :

1. Chaque site $(i, n) \in \mathbb{Z} \times \mathbb{N}$ est *occupable* avec probabilité $1 - p$, et ce indépendamment des autres.

2. Depuis le patch $(i, n) \in \mathbb{Z}\times\mathbb{N}$, il est possible d'atteindre les patchs $(i+i', n+n')$, $i' \in \{-1, 0, 1\}$, $n' \in [\![1, H + 1]\!]$.

Nous pouvons donc nous ramener à un problème de percolation orientée, qui a déjà été étudié dans la littérature. En utilisant un résultat de [HS21], nous pouvons en déduire l'existence d'une probabilité d'extinction critique $p_{crit}(H) \in (0, 1)$ dépendant du paramètre H au delà de laquelle une expansion de population n'est pas possible, quelles que soient les valeurs prises par les autres paramètres.

Theorème 3. *Pour tout $H \in \mathbb{N}$, il existe $p_{crit}(H) \in (0,1)$ tel que pour tout $M \in \mathbb{N}\backslash\{0\}$, $k \in \mathbb{N}\backslash\{0,1\}$, $g \in (0,1)$ et $c \in (0,1/2)$, pour tous $(\xi, h) \in \mathcal{F}_M \times \mathcal{H}_M$ et $p \in (p_{crit}(H), 1]$, si $(\xi^n, h^n)_{n\in\mathbb{N}}$ est le k-parent WFSB de condition initiale (ξ, h) et de paramètres (M, H, g, c, p), alors*

$$\lim_{n\to+\infty} \mathbb{P}\left(\forall (i,j) \in \mathbb{Z} \times [\![1, M]\!], \xi^n_{i,j} \mathbb{1}_{\{h^n_{i,j} \leq H\}} = 0\right) = 1.$$

La preuve repose sur l'utilisation d'un processus appelé processus BOA (pour "Best Occupancy Achievable"), qui encode quels patchs peuvent potentiellement contenir des plantes réelles étant donné la condition initiale et les événements d'extinction, et qui peut être interprété comme un problème de percolation orientée. Ce processus est défini sur l'espace d'états $\mathcal{F}^\infty \times \mathcal{H}^\infty$, où

$$\begin{aligned}\mathcal{F}^\infty &:= \{(O_i)_{i\in\mathbb{Z}} : \forall i \in \mathbb{Z}, O_i \in \{0,1\} \text{ and } \mathrm{Card}\left(\{i \in \mathbb{Z} : O_i = 1\}\right) < +\infty\} \\ \text{and } \mathcal{H}^\infty &:= \{(h_i)_{i\in\mathbb{Z}} : \forall i \in \mathbb{Z}, h_i \in \mathbb{N}\},\end{aligned}$$

de la façon suivante.

Définition 1.3.11. *(Processus BOA) Soit $(O, h) \in \mathcal{F}^\infty \times \mathcal{H}^\infty$. Le processus BOA $(O^{\infty,n}, h^{\infty,n})_{n\in\mathbb{N}}$ de condition initiale (O, h) est la chaîne de Markov à valeurs dans $\mathcal{F}^\infty \times \mathcal{H}^\infty$ telle que $(O^{\infty,0}, h^{\infty,0}) = (O, h)$ et pour tout $n \in \mathbb{N}$, sachant $(O^{\infty,n}, h^{\infty,n})$:*

1. *Soient $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ des variables aléatoires i.i.d suivant une loi de Bernouilli de probabilité de succès p.*
2. *Pour tout $i \in \mathbb{Z}$, si $\mathrm{Ext}_i = 0$ et $O^{\infty,n}_i \mathbb{1}_{\{h^{\infty,n}_i \leq H\}} = 1$, alors nous posons*

$$\begin{aligned} & O^{\infty,n+1}_{i-1} = O^{\infty,n+1}_i = O^{\infty,n+1}_{i+1} = 1 \\ \textit{and} \qquad & h^{\infty,n+1}_{i-1} = h^{\infty,n+1}_i = h^{\infty,n+1}_{i+1} = 0. \end{aligned}$$

   Nous ne faisons rien à cette étape sinon.

3. *Pour tout $i \in \mathbb{Z}$, si $O^{\infty,n+1}_i$ n'a pas déjà été défini durant l'étape 2, nous posons $O^{\infty,n+1}_i = O^{\infty,n}_i$ et $h^{\infty,n+1}_i = h^{\infty,n}_i + 1$.*

Si le processus BOA est construit à partir de la "même" condition initiale (modulo la différence d'espaces d'états) et des mêmes événements d'extinction que le k-parent WFSB, alors pour tout $n \in \mathbb{N}$ et $i \in \mathbb{Z}$:

- Si $O^{\infty,n}_i = 0$, alors le patch i ne peut pas contenir de graines de type 1 viables au début de la génération n.
- Si $O^{\infty,n}_i = 1$, ceci signifie que si le patch i contient des graines de type 1 au début de la génération n, alors celles-ci sont au moins d'âge $h^{\infty,n}_i$. En particulier, si $h^{\infty,n}_i > H$, alors le patch ne peut pas contenir de graines de type 1 viables.

Ainsi, il est possible de construire un couplage entre le k-parent WFSB et le processus BOA de sorte que le k-parent WFSB soit "inclus" dans le processus BOA. L'extinction du processus BOA associé constitue donc une condition suffisante pour l'extinction du k-parent WFSB. Nous pouvons alors utiliser les résultats connus sur le processus BOA via la théorie de la percolation pour en déduire l'existence d'une probabilité d'extinction de patch $p_{crit}(H)$ au-delà de laquelle le k-parent WFSB s'éteint presque sûrement.

L'inclusion du k-parent WFSB dans le processus BOA est généralement stricte, et des déviations peuvent se produire dans l'un des trois cas suivants :

1. Des plantes de type 1 sont bel et bien présentes dans un patch, mais aucune d'entre elles n'est choisie comme parent potentiel.
2. Des graines de type 1 sont présentes dans un patch, mais aucune ne germe.
3. Des graines de type 1 ont intégré la banque de graines il y a moins de $H+1$ générations, mais elles ont toutes déjà germé.

Cependant, lorsque $M \to +\infty$ et $k \to +\infty$, si k croît "plus vite" que M (i.e, s'il existe $\alpha > 1$ tel que $k = \lceil M^\alpha \rceil$), alors la probabilité de chacun de ces événements tend vers 0, et nous pouvons montrer que le k-parent WFSB converge vers le processus BOA.

Theorème 4. *Soit $\alpha > 1$. Pour tout $M \geq 2$, soit $(\xi^{(M),n}, h^{(M),n})_{n\in\mathbb{N}}$ le $\lceil M^\alpha \rceil$-parent WFSB de paramètres (M, H, g, c, p), et soit $(O^{(M),\infty,n}, h^{(M),\infty,n})_{n\in\mathbb{N}}$ le processus BOA associé. Pour tout $M \geq 2$, $i \in \mathbb{Z}$ et $n \in \mathbb{N}$, posons*

$$O_i^{(M),n} := 1 - \prod_{j\in[\![1,M]\!]} \left(1 - \xi_{i,j}^{(M),n}\right)$$

$$h_i^{(M),0} := \begin{cases} \min\{h_{i,j}^{(M),0} : j \in [\![1,M]\!] \text{ et } \xi_{i,j}^{(M),0} = 1\} & \text{if } O_i^{(M),0} = 1 \\ 0 & \text{sinon.} \end{cases}$$

$$\text{et } h_i^{(M),n} := \begin{cases} \min\{h_{i,j}^{(M),n} : j \in [\![1,M]\!] \text{ et } \xi_{i,j}^{(M),n} = 1\} & \text{if } O_i^{(M),n} = 1 \\ h_i^{(M),n-1} + 1 & \text{sinon.} \end{cases}$$

Alors, pour tout $N \in \mathbb{N}$,

$$\mathbb{P}\left(\bigcap_{n=0}^{N} \left(\left\{\forall i \in \mathbb{Z}, O_i^{(M),n} = O_i^{(M),\infty,n}\right\} \cap \left\{\forall i \in \mathbb{Z}, h_i^{(M),n} = h_i^{(M),\infty,n}\right\}\right)\right) \xrightarrow[M\to+\infty]{} 1.$$

Ce résultat de convergence permet de conclure au fait que $p_{crit}(H)$ est bien la probabilité d'extinction critique pour le k-parent WFSB.

1.4 Modèles de métapopulation de type SPOM et banques de graines

1.4.1 Qu'est-ce qu'un SPOM ?

Le résultat de convergence du k-parent WFSB vers le processus BOA présente aussi un intérêt du point de vue de la modélisation des métapopulations. En effet, il permet de faire le lien entre deux familles de modèles de métapopulations : les modèles individu-centrés, dont l'interprétation biologique est plus directe mais l'analyse mathématique théorique généralement difficile, et les modèles de type *Stochastic Patch Occupancy Models*, ou SPOMs. Ces modèles markoviens sont caractérisés par le fait qu'ils n'encodent que la présence ou absence de l'espèce étudiée dans chacun des patchs de la métapopulation, comme le fait le processus BOA. Le processus BOA a ainsi deux interprétations et applications possibles : comme modèle encodant l'effet des événements d'extinction sur la distribution possible de l'espèce, ou comme modèle simplifié de dynamique de métapopulations.

Ne faire dépendre la dynamique que de la présence/absence de l'espèce en négligeant son abondance représente une forte simplification de la dynamique intra-patch, qui a déjà été discutée dans la littérature (voir par exemple [Bag04; DSV03; Han04]). Cette approximation est considérée comme adaptée aux espèces vivant dans des environnements très fragmentés [Han04] tels que les

pieds d'arbre en milieu urbain [DPC11; Oma+19]. Elle rend en contrepartie la collecte de données plus aisée (il est en effet plus facile de collecter des données de présence/absence que des données d'abondance), et permet d'obtenir des résultats théoriques sur ces modèles. Voir par exemple [HO03] pour une présentation de plusieurs de ces résultats. De plus, les modèles SPOMs étant des chaînes de Markov, des estimateurs ont pu être développés pour plusieurs des SPOMs classiques [Moi99; Moi04].

Afin de vérifier la validité de l'approximation sous-jacente aux SPOMs, plusieurs études [AN94; Kee02; OH04] se sont déjà intéressées à comparer la dynamique d'un modèle individu-centré complet à celle d'une approximation de type SPOM, via l'usage de simulations. Leurs conclusions sont que les propriétés de la dynamique du modèle SPOM simplifié approchent généralement bien la dynamique observée dans le modèle individu-centré, pour les régimes de paramètres étudiés. Le résultat de convergence du k-parent WFSB vers le processus BOA complète ces différentes études, en donnant cette fois-ci un résultat théorique plutôt que basé sur des simulations.

Nous allons maintenant présenter brièvement quelques modèles SPOMs classiques. Pour cela, rappelons que l'idée centrale derrière la théorie des métapopulations en environnement fragmenté est le fait que les événements d'extinction locaux (i.e., à l'échelle d'un seul patch) sont fréquents, du fait de perturbations extérieures ou du faible nombre d'individus par patch. Pour autant, l'espèce arrive quand même à se maintenir à l'échelle régionale, grâce à des événements de recolonisation eux aussi fréquents. Le *Propagule Rain Model*, ou PRM [Got91], et le modèle de Levins [Lev69] sont deux modèles SPOMs classiques qui représentent en quelque sorte deux extrêmes en termes de modélisation de la colonisation. En effet, dans le PRM, les nouveaux individus proviennent tous d'une source extérieure (la pluie de propagules qui donne son nom au modèle), alors que dans le modèle de Levins, ils ne peuvent venir que des autres patchs.

La version d'origine du modèle de Levins peut être vue comme un modèle d'îles, au sens où la distance entre les patchs n'intervient pas dans la dynamique de colonisation: les nouveaux individus peuvent provenir d'un patch voisin comme d'un patch lointain avec la même probabilité. Le modèle a depuis été modifié pour intégrer plus de structuration spatiale, et faire dépendre les probabilités de colonisation de la distance entre les patchs (voir par exemple [Moi04]).

Formellement, le PRM et le modèle de Levins structuré spatialement (désigné dans la littérature sous le nom de *spatially realistic Levins model*) sont définis de la façon suivante. Supposons que les localisations des patchs correspondant à l'ensemble discret (généralement fini) $E \subset \mathbb{R}^d$, $d \geq 1$ (le cas le plus fréquent étant $d = 2$), et pour tout $(i, j) \in E^2$, notons $d_{i,j}$ la distance entre les patchs i et j. Le PRM et le modèle de Levins sont chacun définis sur l'espace $\{0, 1\}^E$, et sont caractérisés par les paramètres suivants :

- Dans le cas du PRM, sa *probabilité de colonisation* c et sa *probabilité d'extinction intrinsèque* p.
- Dans le cas du modèle de Levins structuré spatialement, sa *distance de dispersion moyenne* $\delta \in (0, \infty)$, sa *probabilité d'extinction intrinsèque* p et son *paramètre de connectivité* $\gamma \in (0, \infty)$.

Ces deux modèles sont alors définis de la façon suivante.

Définition 1.4.1. *(Propagule Rain Model) Soit $(x_i^0)_{i\in E} \in \{0, 1\}^E$. Le Propagule Rain Model $(X^n)_{n\in\mathbb{N}}$ de condition initiale $X^0 = (x_i^0)_{i\in E}$ et de paramètres (c, p) est la chaîne de Markov à valeurs dans l'espace d'états $\{0, 1\}^E$ dont les probabilités de transition sont les suivantes. Pour tout $n \in \mathbb{N}$, sachant $X^n = (X_{i'}^n)_{i'\in E}$, pour tout $i \in E$,*

- *Si $X_i^n = 0$, alors $X_i^{n+1} = 1$ avec probabilité c, et $X_i^{n+1} = 0$ sinon.*
- *Si $X_i^n = 1$, alors $X_i^{n+1} = 0$ avec probabilité $p(1 - c)$, et $X_i^{n+1} = 1$ sinon.*

En d'autres termes, dans le PRM, à chaque génération, un patch éteint devient occupé avec probabilité c. Un patch occupé s'éteint avec probabilité p, mais est aussitôt recolonisé avec probabilité c. Ainsi, l'extinction est visible avec probabilité $p(1-c)$.

Définition 1.4.2. *(Modèle de Levins structuré spatialement) Soit $(x_i^0)_{i\in E} \in \{0,1\}^E$. Le modèle de Levins structuré spatialement $(X^n)_{n\in\mathbb{N}}$ de condition initiale $X^0 = (x_i^0)_{i\in E}$ et de paramètres (δ, γ, p) est la chaîne de Markov à valeurs dans l'espace d'états $\{0,1\}^E$ dont les probabilités de transition sont les suivantes. Pour tout $n \in \mathbb{N}$, sachant $X^n = (X_{i'}^n)_{i'\in E}$, pour tout $i \in E$,*

- *Si $X_i^n = 0$, alors $X_i^{n+1} = 1$ avec probabilité*

$$C_i^{n+1}(\delta,\gamma) = 1 - \exp\left(-\gamma \sum_{j\in E\setminus\{i\}} X_j^n \exp\left(-\delta^{-1} d_{i,j}\right)\right),$$

  et $X_i^{n+1} = 0$ sinon.

- *Si $X_i^n = 1$, alors $X_i^{n+1} = 0$ avec probabilité $p(1 - C_i^{n+1}(\delta,\gamma))$, et $X_i^{n+1} = 1$ sinon.*

La différence avec le PRM est que cette fois-ci, la colonisation a lieu depuis les autres patchs, et intègre une pondération par la distance. La probabilité de colonisation est ainsi d'autant plus élevée que les patchs voisins sont occupés.

1.4.2 Prise en compte de la banque de graines et inférence

Si les SPOMs ont pu être appliquées avec succès à l'étude de métapopulations d'insectes (voir par exemple [Han11; MSH98]) ou de petits mammifères (voir par exemple [Ozg+06]), leur application à des métapopulations de plantes a longtemps été freinée par la non prise en compte par les modèles SPOMs classiques d'une potentielle banque de graines dans le sol [FW02]. Estimer les paramètres du modèle en négligeant la présence de la banque de graines introduit un biais important sur les taux de colonisation et d'extinction [Fré+13], et augmente le taux d'identification d'un PRM, car la banque de graines laisse une signature proche de celle d'une pluie de propagules [Oma+19]. Plusieurs modèles SPOMs prenant en compte la présence d'une banque de graines ont depuis été développés [Bor+15; Fré+13; Plu+18]. Ces modèles encodent à la fois la présence/absence de plantes et la présence/absence de graines viables, et dans certains cas une structuration en âge de la banque de graines. Il s'agit à chaque fois de modèles de Markov cachés (ou *Hidden Markov Models*, abrégé en HMM) : la présence/absence de graines est considérée comme un état caché non visible par l'observateur, qui influence la présence/absence de plantes dans le patch (l'état observé) et peut être inféré à partir des observations. C'est d'ailleurs aussi le cas du processus BOA : l'âge des graines viables les plus jeunes dans la banque de graines d'un patch est un état caché, duquel dépend la possibilité d'observer ou non des plantes dans le patch.

Le modèle introduit dans [Plu+18] est une variante avec banque de graines du PRM. L'estimateur associé repose sur une variante de l'algorithme EM [DLR77] pour les HMMs. Il a pu être appliqué avec succès à des données réelles de présence/absence d'adventices dans des champs [Plu+18] et dans des vignes [Kaz+21]. En supposant que les localisations des patchs sont toujours données par $E \subset \mathbb{R}^d$, cette variante du PRM est cette fois-ci définie sur l'espace d'états $\{(0,0),(1,0),(1,1)\}^E$. Si $(z_i)_{i\in E} = (s_i, x_i)_{i\in I} \in \{(0,0),(1,0),(1,1)\}^E$ correspond à l'état de la métapopulation durant la génération $n \in \mathbb{N}$, alors pour tout $i \in E$, $s_i = 1$ si la banque de graines du patch i contient des graines (viables) au début de la génération n, et $x_i = 1$ si le patch i contient des plantes durant la génération n. Dans le PRM avec banque de graines, la phase de colonisation se produit après la phase de germination, et il n'est donc pas possible d'avoir $s_i = 0$ et $x_i = 1$.

Le modèle est caractérisé par trois paramètres :

- une probabilité de colonisation, notée c;
- une probabilité de germination (et de survie des plantules jusqu'à l'âge adulte), notée g;
- une probabilité de mort de la banque de graines (conditionnellement à la non-germination, car les plantes produisent toujours de nouvelles graines), notée d.

Formellement, il est défini de la façon suivante.

Definition 1.4.3. *(PRM avec banque de graines) Soit $(z_i^0)_{i\in E} \in \{(0,0),(1,0),(1,1)\}^E$. Le PRM avec banque de graines $(Z^n)_{n\in\mathbb{N}}$ de condition initiale $Z^0 = (z_i^0)_{i\in E}$ et de paramètres (c,g,d) est la chaîne de Markov à valeurs dans l'espace d'états $\{(0,0),(1,0),(1,1)\}^E$ dont les probabilités de transition sont les suivantes. Pour tout $n \in \mathbb{N}$ et $i \in E$, sachant $Z^n = (Z_i^n)_{i\in E}$,*

- *Si $Z_{i'}^n = (0,0)$, alors*

$$Z_{i'}^{n+1} = \begin{cases} (0,0) & \text{avec probabilité } 1-c, \\ (1,0) & \text{avec probabilité } c(1-g), \\ (1,1) & \text{avec probabilité } cg. \end{cases}$$

- *Si $Z_{i'}^n = (1,0)$, alors*

$$Z_{i'}^{n+1} = \begin{cases} (0,0) & \text{avec probabilité } d(1-c), \\ (1,0) & \text{avec probabilité } (1-d(1-c))(1-g), \\ (1,1) & \text{avec probabilité } (1-d(1-c))g. \end{cases}$$

- *Si $Z_{i'}^n = (1,1)$, alors*

$$Z_{i'}^{n+1} = \begin{cases} (1,0) & \text{avec probabilité } 1-g, \\ (1,1) & \text{avec probabilité } g. \end{cases}$$

L'interprétation de cette définition est la suivante. Si un patch ne contient pas de graines, alors il est colonisé avec probabilité c, et les graines correspondantes germent avec probabilité g. Si le patch contient des graines, ou bien celles-ci germent (avec probabilité g), donnant des plantes qui vont produire de nouvelles graines, ou bien aucune ne germe. La banque de graines peut alors s'éteindre avec probabilité d, extinction aussitôt compensée par une recolonisation avec probabilité c. Voir la Figure 1.14 pour une illustration de la dynamique.

Durant ma thèse, j'ai cherché à appliquer l'estimateur associé au jeu de données des plantes dans les pieds d'arbre du 12ème arrondissement de Paris décrit plus haut (voir la Section 1.1.2), afin de confirmer les résultats de [Oma+19] relatifs à la présence d'une banque de graines dans les pieds d'arbre. Utiliser directement l'estimateur introduit dans [Plu+18] n'a cependant pas été possible, car la quantité de données récoltées n'était pas suffisante pour pouvoir différentier suffisamment souvent le modèle avec banque de graines du modèle sans banque de graines. Le jeu de données représente pourtant 10 ans de relevés, dans des rues allant de 30 à 200 pieds d'arbres. Ce problème est donc susceptible de se poser avec d'autres jeux de données comprenant moins d'observations.

L'introduction d'une nouvelle métrique, appelée *Seed Bank Characteristic Event probability*, ou probabilité SBCE, a permis de pallier ce problème. Cette métrique, présentée en détail dans le Chapitre 5, mesure la contribution de la potentielle banque de graines à la dynamique observée. Elle est égale à zéro en l'absence de banque de graines (i.e, lorsque $d = 1$), mais aussi lorsque la banque de graines n'a pas d'effet visible sur la dynamique observée (i.e, lorsque $c = 1$ et/ou $g = 1$). De plus, lorsqu'une banque de graines est présente, la probabilité SBCE est d'autant plus faible que les événements de colonisation sont fréquents et les événements d'extinction rares, situations

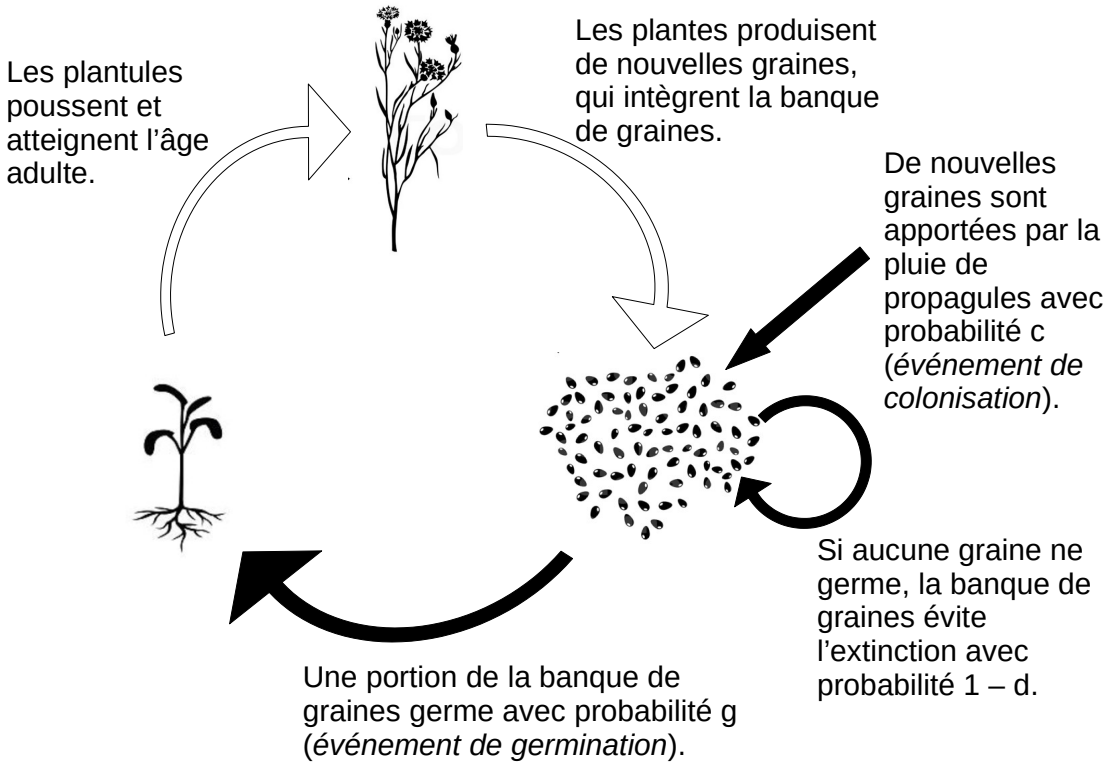


Figure 1.14: Illustration de la dynamique du PRM avec banque de graines.

pour lesquelles la présence de la banque de graines n'influe que peu la dynamique et donc pour lesquelles conclure à la présence/absence d'une banque de graines à partir des observations de plantes est le plus difficile. Concrètement, la probabilité SBCE correspond à la probabilité que les plantes présentes dans un patch produisent des graines qui germeront de façon différée, sans qu'un événement de colonisation ne se produise en même temps. Ainsi,

$$\mathbb{P}_{SBCE} = g\frac{(1-g)(1-c)(1-d)}{1-(1-g)(1-c)(1-d)}.$$

Le but principal du Chapitre 5 est de montrer que même si cette métrique est moins informative que de savoir si une banque de graines est effectivement présente, elle nécessite en pratique moins de données pour être suffisamment précise pour les applications envisagées. De plus, elle est moins sensible aux erreurs de relevés, ce qui la rend applicable à des jeux de données issus par exemple des sciences participatives. En fonction de la question écologique d'intérêt, la probabilité SBCE peut donc représenter une métrique intéressante.

D'un point de vue théorique, l'approche utilisée illustre l'intérêt de s'intéresser à la détectabilité de la banque de graines, en particulier dans des modèles de type SPOM, et de prendre en compte cet aspect pour mesurer les performances d'estimateurs. L'application de la probabilité SBCE au jeu de données des plantes dans des pieds d'arbres a de plus permis de mettre en évidence une contribution significative de la banque de graines à la dynamique observée pour certaines espèces végétales. La méthodologie utilisée peut aussi être appliquée à d'autres jeux de données.

1.5 Perspectives

Pendant ma thèse, j'ai donc développé deux familles de modèles de génétique des populations adaptés à l'étude de populations en expansion :

- une famille de modèles continus en espace et en temps, basés sur le processus Λ-Fleming Viot spatial,
- une famille de modèles discrets en espace et en temps, adaptés à la modélisation de métapopulations de plantes dans un environnement fragmenté et perturbé.

Mes travaux de thèse portent principalement sur la construction de ces modèles, et sur le développement d'outils de reconstruction des généalogies d'échantillons d'individus. La continuation logique de ces travaux est l'utilisation de ces modèles pour étudier la diversité génétique dans des populations en expansion, en particulier au niveau du front. Ceci passe entre autres par le développement de techniques permettant d'exploiter les processus duaux, qui encodent les généalogies.

Pour cela, j'envisage d'explorer les deux pistes suivantes. Tout d'abord, il est possible de s'intéresser aux généalogies d'individus *conditionnés à être réels*, c'est à dire d'individus observables. En effet, si un individu est réel, alors au moins l'un de ses ancêtres potentiels dans la condition initiale est réel lui aussi. De plus, il existe une lignée ancestrale composée uniquement d'individus réels, qui remonte vers le (seul) vrai ancêtre de l'individu dans la condition initiale. En l'absence de mutations, cet ancêtre porte le même type marqué que l'individu considéré.

Plutôt que de chercher l'ensemble des ancêtres potentiels d'un individu réel, nous pouvons nous focaliser sur l'étude de cette lignée ancestrale spécifique, qui vérifie des propriétés particulières. Elle est ainsi conditionnée à remonter jusqu'à la zone de laquelle est partie l'expansion, la seule à contenir initialement des individus réels. De plus, dans le cas du k-parent WFSB, elle est conditionnée à éviter les événements d'extinction et à ne rester dormante que durant au plus H générations consécutives. L'étude théorique porterait sur la distribution de la "vraie" lignée ancestrale d'un individu réel, et sur la façon dont plusieurs lignées associées à des individus différents interagissent. Les résultats obtenus pourraient être utilisés pour simuler ces lignées, ou pour construire des estimateurs permettant de faire de l'inférence à partir de données réelles.

Puis, une autre approche possible consiste à étudier le modèle au moyen de simulations, approche qui se heurte à un obstacle de taille : la façon intuitive de coder le processus n'est pas très efficace d'un point de vue computationnel. Ceci est particulièrement vrai pour le k-parent SLFV et le ∞-parent SLFV, du fait de la construction basée sur un processus ponctuel de Poisson. Simuler le processus de façon plus efficace nécessiterait de s'intéresser à des constructions ou représentations alternatives du processus, ou d'exploiter des résultats théoriques sur la dynamique du processus. Pouvoir simuler plus efficacement le ∞-parent SLFV pourrait entre autres permettre d'affiner la conjecture selon laquelle le modèle appartient à la classe d'équivalence de l'équation KPZ, en plus de permettre d'étudier l'apparition de secteurs dans le front d'une population en expansion.

Même si ma thèse porte principalement sur des modèles de génétique des populations, elle contient aussi des contributions à la modélisation et l'étude de métapopulations. En particulier, l'un des modèles limites obtenus, le processus BOA, peut être interprété comme un modèle de type SPOM. Il s'agit d'une version simplifiée du modèle de Levins structuré spatialement, qui intègre de plus la présence potentielle d'une banque de graines. Il complète donc la variété de modèles de type SPOM déjà existants. Des travaux en cours, en collaboration avec Nathalie Machon et Amandine Véber, portent sur le développement d'un estimateur associé au processus BOA, qui permettrait de faire de l'inférence à partir de données de présence/absence de plantes. La méthode d'estimation utilise les techniques classiques pour les modèles de Markov cachés, appliquées ici au cas particulier du processus BOA. Nous utiliserons l'estimateur en association avec celui associé au Propagule Rain Model (ou PRM) avec banque de graines, et l'appliquerons au jeu de données

des plantes dans les pieds d'arbres d'alignement. L'objectif est d'identifier les points d'entrée de nouvelles graines, et les zones correspondant à des corridors écologiques. La méthode utilisée pourrait être appliquée de façon plus générale à des métapopulations de plantes dans lesquelles la présence d'une banque de graines est soupçonnée.

1.6 Structure de la thèse

Ma thèse est structurée en deux parties. La première, composée des Chapitres 3 et 4, s'intéresse aux modèles de populations en expansion basés sur le processus Λ-Fleming Viot spatial, ou SLFV. Le Chapitre 3 porte sur l'intégration du concept d'"individus fantômes" au Λ-Fleming Viot spatial. Le processus obtenu est appelé k-parent SLFV, le paramètre k faisant référence au nombre de parents potentiels choisis. Le Chapitre 3 se focalise ensuite sur la construction de la limite $k \to +\infty$ de ce processus, limite désignée sous le nom de ∞-parent SLFV. Le chapitre est issu d'un article soumis à *ESAIM - Probability and Statistics*. Le Chapitre 4 s'intéresse quant à lui aux propriétés de croissance du ∞-parent SLFV, et montre que l'expansion de la zone occupée par les individus réels est linéaire en temps. Afin de comprendre comment la dynamique caractéristique observée au niveau du front d'expansion peut conduire à une expansion plus rapide qu'initialement conjecturé, il étudie de plus un modèle jouet, pour lequel il est possible d'obtenir une expression explicite pour la vitesse de croissance. Ce chapitre est issu d'un article en collaboration avec Amandine Véber, qui a été soumis à *Electronic Journal of Probability*.

La deuxième partie de ma thèse, plus appliquée, porte sur le rôle joué par la banque de graines dans la dynamique de métapopulations de plantes vivant dans un environnement fragmenté et soumis à des événements d'extinction locaux fréquents. L'objectif est d'utiliser les résultats obtenus pour mieux comprendre la dynamique des plantes dans les pieds d'arbres d'alignement en milieu urbain. Le Chapitre 5 est une étude préliminaire menée sur le jeu de données d'observations de plantes dans les pieds d'arbres de Paris. Il s'agit d'un travail en collaboration avec Nathalie Machon, Jean-Baptiste Mihoub et Alexandre Robert, qui a été publié dans *Methods in Ecology and Evolution*. Enfin, le Chapitre 6 porte sur la construction et l'étude d'un modèle théorique adapté, basé sur une variante avec individus fantômes du modèle de Wright-Fisher avec banque de graines. Le résultat principal de ce chapitre est l'existence d'une probabilité critique d'extinction de patch dépendant des paramètres de banque de graines au delà de laquelle une expansion de population n'est pas possible. L'article correspondant a été accepté dans *Theoretical Population Biology*.

Chapter 2

Introduction (in English)

This chapter corresponds to a translation of part of Chapter 1 *(most of Section* 1*, and Sections* 3 *to* 6*). The parts which were not translated are presentations of classical populations genetics models or stochastic growth models:*

- *The Eden growth model. See [Ede61] for the original definition of the model, and [WLB95] for a continuous in space version.*
- *The Wright-Fisher model, its large population limit and the associated dual process. See [Eth11] for a detailed presentation.*
- *Variants of the Wright-Fisher model incorporating with selection, fluctuating population sizes, a spatial structure, or a seed bank component. Most of them can be found in [Eth11], except variants with a seed bank component : see [KKL01], [Bla+16] and references therein.*
- *The spatial* Λ*-Fleming Viot process and its variant with selection. See e.g [BEV10], [EFP17] and [EVY20].*

2.1 Biological motivations

Population expansions are a common phenomenon that occurs at temporal, spatial and biological scales. Examples are varied: invasive species, tumors, but also some tree species in Europe, or successive expansions of new variants of a same virus. Moreover, many species whose geographic distribution can currently be considered as stable have in fact been expanding in the past, following a change in climate, the species introduction in a new environment, or the emergence of new mutations. For instance, the plant species *Veronica persica* (see Figure 2.1), characterized by small blue and white flowers, can be found everywhere in France [MO22], and can often be observed in parks and gardens. However, this species actually originated from southwestern Asia. It was introduced all around the globe, and is considered as invasive in many countries. On a longer timescale, many plant and animal species seen everyday in Europe underwent an expansion at the end of the last ice age, and have (re)colonized all or part of Europe. Moreover, some species of trees [SNS08; SS07], grass [Van+07], reptiles or amphibians [Ara+08; APR05] have not finished, and are still expanding.

Population expansions leave a imprint on observed genetic diversity patterns, which stays visible even after the expansion is over. Indeed, two individuals are genetically closer when they have a recent common ancestors. Moreover, if the individuals are haploid (that is, generally, if they only have one parent), then the differences between two individuals having a common ancestor can only be caused by mutations having occurred since this ancestor. Therefore, the observed genetic diversity is directly linked to the shape of genealogies. However, a population expansion yields

Figure 2.1: *Veronica persica*

different genealogies compared to those associated with populations at equilibrium. Indeed, when we go back in time and reconstruct the genealogies of a sample of individuals, their ancestors are closer and closer to the area from where the expansion started. In addition, individuals near the front edge, where population densities are lower for a same amount of resources, face less competition and can reproduce more.

Other events, such as extinction events, also affect genealogies, and hence the observed genetic diversity contains information about the population's past. Exploiting this information can give insight on how species have adapted to past climate changes, or on human evolutionary history [Hew00]. In a completely different field of application, analyzing patterns of genetic diversity in tumors may allow us to learn more about the events that occur early in the growth phase, when the tumor is too small to be detected and analyzed. Therefore, one important question in population genetics is how to develop tools to extract information from the observed genetic diversity and reconstruct the past of a population, which is more difficult in the case of expanding populations.

The factors that trigger a population expansion are varied: the introduction of a species in a new environment in which it can thrive, a modification of the environment or climate, or the appearance of a new mutation that makes the individuals that carry it more competitive. However, understanding how to trigger or prevent a population expansion in practice is not always easy, as shown by the failure of some programs for the reintroduction of threatened or extinct species or (conversely) for the eradication of invasive species. To illustrate this, consider the example of a deleterious mutation (i.e., a mutation leading to a decrease in the number of descendants of the individuals carrying it). In a population at equilibrium, the individuals carrying this mutation reproduce less and are in competition with more adapted individuals. The sub-population of individuals carrying the deleterious mutation is bound to become extinct quickly, even if no measures are taken to control or prevent its expansion. However, this is not always the case in an expanding population. Indeed, the mutation can persist over longer timescales by "surfing" on the front, where population densities are lower and competition less strong (but where genetic drift is stronger). Furthermore, if the mutation increases the dispersal ability, it is possible under certain conditions for the mutation to completely invade the front, despite its selective disadvantage ("survival of the fastest", [CP16; Def+19]). In the long run, the corresponding sub-population will eventually become extinct, but not before the expansion is over and the entire possible range occupied. In addition, the expansion of the sub-population of individuals not carrying the mutation is slowed. This surfing phenomenon of deleterious mutations at the front has been observed experimentally [Dit+13; Def+19] as well as in real populations, such as cane toads in Australia [Hud+15] (see also [Def+19] for other examples).

The example of deleterious mutations in an expanding population, which can also be generalized to the case of a species with a selective disadvantage compared to another species with which it is in competition, is an example of what counter-intuitive phenomena can occur during a population expansion and prevent us from understanding what favors or prevents the expansion of a (sub)population. However, answering this question is crucial to understand how to control or

eradicate invasive species, which represent a threat to biodiversity [BCB16; BBR19], ecosystems [Kum+15; WCV16], economy [Bra+16; Dia+21], agriculture [Pai+16] and/or health [She+11]. On a completely different biological scale, tumors can also be considered as expanding populations, and understanding how to prevent their growth could lead to the development of new therapies. Furthermore, understanding what *prevents* a population expansion may also give insight on how to design efficient species reintroduction or protection programs, or on how to increase connectivity in fragmented environments. During my PhD, I focused on two biological questions of interest related to expanding populations, which I will now present.

2.1.1 Genetic diversity at the front of an expanding population

The first question is motivated by an experiment of Oskar Hallatschek and collaborators [Hal+07; HN10]. This experiment consists in letting fluorescent strains of bacteria (*E. coli*) or yeast (*S. cerevisiae*) grow in Petri dishes. The fluorescence color is encoded by a single gene, and is the only difference between the two strains. Moreover, it is selectively neutral, and does not influence the behavior of the bacteria. In other words, it is simply a marker, allowing to visualize the types distribution.

The results obtained are the following ones. While the two fluorescent strains are well-mixed in the area initially occupied, this is not the case at all at the front edge. Indeed, the front is divided into sectors in which all individuals have the same fluorescence color. The sectors are also relatively large and have very well-defined boundaries. The shape and the number of sectors depends on the species considered, as well as on the initial shape of the colony (line or disk).

The emergence of these sectors is due to a combination of two factors: lower population densities at the front, leading to a strong genetic drift (i.e., strong stochastic fluctuations in the frequencies of the different types), and a higher reproduction rate in this area (due to less competition) which amplifies the effect of genetic drift. In dimension 1, due to genetic drift, one of the two types can become extinct at the front. This type is always present where the population densities are higher, but in order to "catch up" with the front, it is now in competition with a higher number of individuals of the other type: its neighbors and the individuals forming the front. With high probability, the "updated" front will only be composed of individuals of the other type, which again will be able to contribute more easily to the next advance of the front. See Figure 2.2 for an illustration of the phenomenon, which can be considered as successive founder effects.

Therefore, in dimension 1, population expansions lead to a decrease in genetic diversity in the expansion direction, due to the stochasticity in reproduction. The resulting gradient of genetic diversity has been mainly studied by means of simulations [EFP09; HN08], and more recently analytically [DF16]. This decay of genetic diversity in the expansion direction can be observed in real populations (see for example [Hew96; Mac+96; Ros+02; Ram+05]), and can be used to reconstruct past expansions (see for example [Hew00; Tem02]). In dimension 2 or in larger dimensions, the same phenomenon occurs in each expansion direction. However, the type which goes extinct in different directions is not necessarily the same. The emergence of the sectors is then due to correlations between which genetic types become extinct in different but close directions.

The emergence of such sectors in expanding populations has also been observed in real populations of tortoises [Gra+13] and in tumors [Sot+15]. However, the presence of areas in which all individuals have the same type is not necessarily the sign of a past or ongoing expansion, and may simply be due to spatially variable natural selection. This is an obstacle to the use of genetic data to reconstruct past expansions, but also to the identification of selectively advantageous mutations [Cur+06; Sch+07]. In the case of an expansion combined with natural selection, as explained above, deleterious mutations can also form sectors at the front edge [HN10]. This is of course also the case for selectively advantaged mutations, but the corresponding sectors are characterized by a differ-

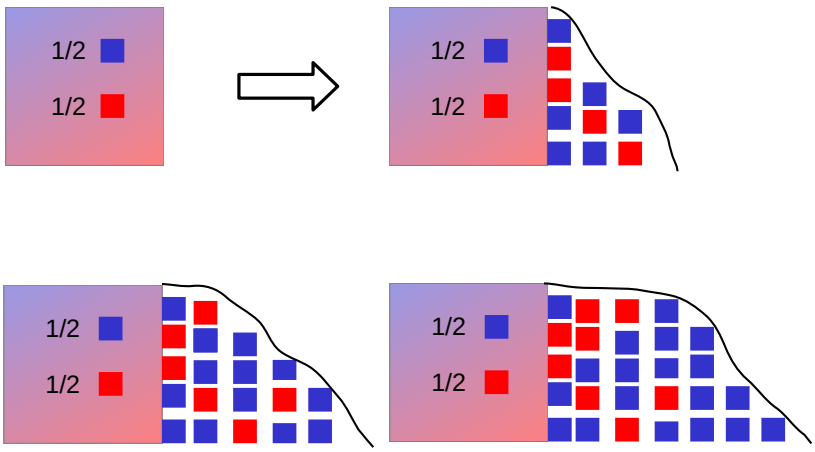


Figure 2.2: Decreasing genetic diversity at the front of an expanding population. A large number of individuals are present in the area from which the expansion starts. The two possible types (blue ou red) are selectively neutral and initially present in the same proportions.

ent opening angle than the sectors corresponding to the spread of neutral or deleterious mutations [HN10].

In order to develop tools allowing to disentangle the effect of an expansion from the one of natural selection, [Hal+07] investigates the morphological properties of boundaries separating sectors. This article shows that these boundaries behave as superdiffusive random walks, contrary what happens in the case of a selectively advantageous mutation (when there is no expansion). This distinctive feature may potentially be used to distinguish between expansion and selection. In the case of an expansion combined with selection, [HN10] studies a simple model of random walks modeling the boundaries between sectors and merging upon collision. The article shows that it may be possible to use the opening angles of the sectors to distinguish neutral mutations from selectively advantageous ones. From a biological viewpoint, two random walks merging corresponds to the disappearance of the corresponding sector.

Therefore, how sectors emerge is well understood qualitatively, and can be reproduced by means of simulations [ELC04; HN08; KCE06; EFP09]. However, there is a lack of quantitative results regarding these sectors. The approach used in [Hal+07; HN10] only allows to study the "post-formation of sectors" regime. From a theoretical viewpoint, the main results that do not presuppose the existence of sectors are restricted populations living in a one-dimensional space. In dimensions 2 and 3, [Dur18] considers a model similar to the one studied in [HN08], and studies in a relatively informal way how sectors emerge in tumors, by considering the genealogies of a sample of cells. Although the study does show the existence of sectors, they are too small to be observed experimentally by biopsy.

Therefore, the emergence of sectors in the front of expanding populations is relatively poorly understood theoretically, which prevents the use of these sectors to perform statistical inference on observed genetic diversity. One of the main reasons for this lack of theoretical results is simple: the lack of stochastic population genetics models adapted to the study of expanding populations (but see for instance [CM07; DF16; JMW12]). Indeed, classical models of expanding populations generally have at least one of the following three limitations:

- they are deterministic, while we have seen above that the emergence of sectors is stochastic

by nature [HN08];

- they are only defined in dimension 1;
- they are not associated to tools allowing to study genetic diversity.

Population genetics models, on the other hand, do allow to study the evolution of genetic diversity, by means of dual processes encoding genealogies. However, in order to construct these dual processes, it is often necessary to assume that population densities are constant, which is not the case in expanding populations.

The main goal of my PhD was to develop population genetic models with a spatial structure, adapted to expanding populations and allowing the study of genetic diversity patterns at the expansion front. These models are a prerequisite to study the emergence and properties of the sectors experimentally observed. The approach I used is based on the adaptation of the technique introduced in [DF16] and [HN08], in order to integrate it to several population genetics models. Originating from interacting particle systems theory, it is based on filling empty areas with "ghost" individuals, which can reproduce, but with a strong selective disadvantage against real individuals. The reproduction of ghost individuals models stochastic fluctuations in population densities, and makes it possible not to predetermine the expansion dynamics. See Section 2.2 for a more detailed presentation of this approach, and of how I applied it to the spatial Λ-Fleming Viot process and to a variant of the Wright-Fisher model.

2.1.2 Spread of plant species in a urban environment along urban tree bases

The second question is related to the dynamics of wild herbaceous plants living in a very specific environment: urban tree bases along streets in cities (see Figure 2.3 for an illustration). These tree bases are very small (around one square meter), and are they are regularly trampled, weeded,... so local extinction events are frequent.

Figure 2.3: (a) Urban tree base near Bordeaux, France. (b) Urban tree base in Bayonne, France. (c) Street with urban tree bases near Bayonne, France.

The interest of this system is twofold: ecological and theoretical. From an ecological viewpoint, tree bases can potentially form ecological corridors between larger green spaces, such as parks and gardens, and contribute to overall connectivity, which improves the overall quality of the urban ecosystem. However, this does not necessarily occur case in practice, due to frequent disturbances leading to more or less localized extinctions. Therefore, it is worth investigating to what extent can urban tree bases actually serve as ecological corridors, and if so, for which species. For species for which this is not the case, a theoretical study can help understanding how to modify tree bases management methods in order to make this possible.

Moreover, from a theoretical viewpoint, urban tree bases are highly suited to mathematical modeling. Indeed, they are well-delimited, of similar sizes, along a line, and often equidistant. As they are regularly weeded by gardeners, the plants they contain can generally not live more than one

year, and generations can be considered as non-overlapping. All these characteristics correspond to assumptions that are often necessary to perform an analytical study of a populations dynamics model, and which are indeed verified here.

Several studies have shown that tree bases in some streets of the cities of Paris [Oma+19] or Montpellier [DPC11] were indeed used as ecological corridors by some plant species. These plants escape from the surrounding green spaces (parks, gardens,...) as seeds, and colonize more or less quickly the neighboring streets generation after generation. This could be a double-edged sword: indeed, because of the frequent disturbances they undergo, urban tree bases can also potentially facilitate the spread of invasive plant species, as it is the case for railroad tracks and *Senecio inaequidens* in Paris [Bla+15a], for instance. More generally, as it highly disturbed, the urban environment is rather favorable to some exotic species (i.e. species introduced deliberately or not due to human activity) [Mur+07; Pyš98], of which invasive species are a particularly dynamic subset. This is at least partly related to the presence of parks and gardens, which are known to contain a large number of exotic species and to potentially act as reservoirs for seedlings [Deh+07a; Deh+07b; Smi+06], and is even more true when urbanization is strong [Kow95; MPS00; MP90; Mur+07]. In her PhD thesis, Noélie Maurel [Mau10] studied the distribution of plant species in Paris region, and showed that more invasive species can be found in urban environments (mainly cities, small parks and former industrial sites) and agricultural environments compared to forests or open environments.

However, while tree bases do harbor a large number of exotic species, they actually contain relatively few invasive species, which is surprising at first glance given the highly disturbed nature of this environment. The study conducted in [Che+08] provides possible answers: invasive species are generally species characterized (among other things) by strong dispersion [Sak+01; WG04]. However, since tree bases are a highly fragmented environment, dispersal increases the probability of seeds falling outside tree bases, where they cannot germinate. Therefore, there may be selection against dispersal, as it has for instance been observed for *Crepis sancta* [Che+08]. Species that are invasive in other environments may not be able to survive in urban tree bases, which illustrates results from [ABH00] regarding how the characteristics that make a species invasive are highly ecosystem-dependent.

If the study conducted in [Che+08] allows us to understand why invasive species are less represented among plants found in urban tree bases, it also raises questions. Indeed, if dispersal is counter-selected and if extinction events are frequent, how do plant species manage not to go extinct? One possibility is the continuous supply of new seeds from an external source, such as parks and gardens. Another possibility is the presence of a seed bank in the soil, i.e. viable but dormant seeds that can potentially germinate several generations after their production. Indeed, the study conducted in [Oma+19] suggests an influence of a seed bank on the dynamics of some plant species growing in tree bases. Assessing whether this is true in practice, for instance by taking soil samples and placing them into germination chambers, is however difficult to implement. This motivates the construction and study of mathematical models, in order to carry out theoretical analyses and statistical inference using real data.

Seed bank models can broadly be divided in two categories: models from population genetics, whose goal is to study the effect of a seed bank on genetic diversity and to make inference from genetic data, and population dynamics models, which focus on the evolution of the number of individuals or the presence/absence of the species. In both cases, these models are relatively recent.

Regarding models from population genetics, most models are based on the one introduced in [KKL01]. This article adds a seed bank to the Wright-Fisher model, which is a classical population genetics model. In the original model of [KKL01], how long can seeds stay dormant without loosing viability is bounded, but this assumption has since been relaxed [Bla+13; Bla+16]. The models of

[Bla+16] and [KKL01] have also been used to construct more general seed bank models incorporating selection [Koo+17] or a spatial structure [HP17; HN21; GHO22]. In some cases, inference can be made from genetic data [Bla+20; Sel+20; Tel+11], allowing to detect the presence of a seed bank and/or to take its presence into account when estimating other parameters.

While models from population genetics often assume constant population sizes, this is not the case in models from population dynamics, which model the evolution of the number of individuals [Jar+95; LCP19] or the presence/absence of the species in each patch [AP01; Bor+15; FW02; Plu+18] (see [LCP19] for a more complete list of references). I was more specifically interested in SPOM-type models (*Stochastic Patch Occupancy Models*, see Section 2.3.1), which are particularly well-suited to the study of urban tree bases [DPC11; Oma+19] and allow for parameter inference in some cases [Kaz+21; Plu+18].

During my PhD, I investigated how the potential presence of seed bank would affect plant dynamics in urban tree bases, and to what extent the ability to form a seed bank could be selected in this type of environment. I investigated this question from two angles:

- A more theoretical angle, involving the development of a model for plant dynamics in urban tree bases that incorporates a seed bank.

This model is based on ideas from several different models, in order to account for the potential presence of a seed bank [Bla+16; KKL01] and to feature local extinction events [DF16; HN08]. This model is more generally suited to plant metapopulations living in a fragmented environment characterized by frequent local extinction events. I used this model to show the existence of a critical patch extinction probability above which a seed bank is needed to survive and successfully colonize other patchs (using an argument from percolation theory), and to provide a theoretical validation for a simplified SPOM-type population dynamics model that is easier to study. However, it can also be used to study genetic diversity in this kind of metapopulation (see Section 2.2.3). See Section 2.2.3 and Chapter 6.

- A more practical angle, based on inference from real data.

I used a dataset of floristic inventories carried out annually from 2009 to 2018 on 1324 tree stands in Paris by Nathalie Machon's team [Mac20]. Using the model and estimator introduced in [Plu+18], I introduced a metric that measures the contribution of a potential seed bank to the observed plant dynamics. Applying this metric to the dataset confirmed results from [Oma+19]. Furthermore, the study of the performances of the associated estimator showed that this metric can be applied to a wider variety of datasets, such as datasets from citizen science. See Section 2.3.2 for a presentation of the metric, and see Chapter 5 for the definition and study of this metric.

2.2 Some population genetics models for expanding populations

As most other population genetics models, the Wright-Fisher model and the spatial Λ-Fleming Viot process assume that population sizes are constant. As explained above, this assumption ensures the existence of a dual process encoding genealogies. Moreover, in the case of the Wright-Fisher model, it is possible to relax this condition, provided that the expansion (or fluctuation) scenario is defined in advance. The expansion dynamics are then fixed, and stochastic fluctuations are predetermined rather than directly generated by the stochasticity in reproduction dynamics. This motivates the construction and study of new population genetics models for expanding populations, in which fluctuations in population sizes are a consequence of the reproduction dynamics, while still being associated to a dual process.

In this section, I will start by presenting the approach introduced in [DF16; HN08]. This approach is based on interacting particle systems theory, and yielded first theoretical results on genetic diversity at the front of expanding populations. I will simultaneously explain how to study genetic diversity in this type of model, and I will highlight to what extent the results and techniques presented can be generalized to other models. Then, I will explain how I applied this idea to the spatial Λ-Fleming Viot process and to a variant of the Wright-Fisher model with a seed bank component.

2.2.1 Ghost individuals

In [DF16] and [HN08], the authors consider demes along a line and indexed by $L_n^{-1}\mathbb{Z}$, $L_n > 0$. Each deme contains exactly $M_n \in \mathbb{N}\backslash\{0\}$ individuals, each of type 1 or type 0. From a modeling viewpoint, the type 0 individuals are "ghost" individuals, and correspond in fact to empty locations, while type 1 individuals correspond to real and observable individuals.

In the model from [DF16], which is a variant of the one initially introduced in [HN08], the reproduction dynamics is based on a model coming from the theory of interacting particle systems: the biased voter model (here in favor of type 1 individuals). In the voter model, during a voting event involving the ordered pair $(x, y) \in (L_n^{-1}\mathbb{Z} \times [\![1, M_n]\!])^2$, the individual located in y "convinces" the individual in x, who takes its type. The bias in favor of type 1 individuals is reflected in the fact that type 1 individuals convince other individuals at a higher rate than type 0 individuals. In the context of the modeling of expanding populations, voting events are interpreted as reproduction events in the following way. If the event involves the pair (x, y), then the individual in x dies and is replaced by a descendant of the individual in y.

The model dynamics are defined as follows. First, let $\xi_t^n(x)$ represent the type of the cell living in x at time t. To each ordered pair (x, y) of interacting individuals (here, located in neighboring patches), we associate two i.i.d. Poisson point processes on $\mathbb{R}_+$: a process $P^{n,(x,y)}$ with intensity $r_n > 0$, and a process $\widetilde{P}^{n,(x,y)}$ with intensity θR_n^{-1}. Then,

- At each jump of $P^{n,(x,y)}$, the individual in x dies, and is replaced by a descendant of the individual in y. In other words, we set

$$\xi_t^n(x) = \xi_{t-}^n(y).$$

- At each jump of $\widetilde{P}^{n,(x,y)}$, the same thing occurs, but *only if* y *is of type* 1. Therefore, we set

$$\xi_t^n(x) = \xi_{t-}^n(y) + (1 - \xi_{t-}^n(y))\xi_{t-}^n(x).$$

In order to add genetic diversity, it is possible to use the concept of *tracer dynamics* from [DF16; HN08]. Rather than adding different types and following the density of each, the idea is to "mark" some type 1 individuals with a neutral mark (which is transmitted to one's descendants), and follow the evolution of the distribution of individuals bearing this mark. To do so, we introduce the notation $\eta_t^n(x)$ to encode whether the cell living in x at time t is marked. In other words, $\eta_t^n(x) = 1$ if this cell is of type 1 and marked, and $\eta_t^n(x) = 0$ otherwise.

In order to construct the dual process, observe that whenever a reproduction event affects the pair (x, y), if given by the Poisson point process $\widetilde{P}^{n,(x,y)}$, then the individual in x takes the type of the individual in y at time $t-$ *if, and only if* the latter is of type 1, and keeps the same type otherwise. In other words, the individual in x at time t is of type 0 if, and only if, the individuals in x and y at time $t-$ are both of type 0. Therefore, it is possible to consider the individuals living in x and y at time $t-$ as the two "potential parents" of the individual in x at time t. Technically, if the individual in y is of type 0, then there is no reproduction, and the individuals in x at time $t-$ and t are in fact the same. However, I will use this terminology in order to fit to the terminology used regarding population genetics models with selection.

Because of the use of ghost individuals and tracers, the model introduced in [DF16] is fundamentally different from classical population genetics models with selection. The main difference comes from the way the type of an individual can be determined given the genealogy. For instance, in the Wright-Fisher model with selection, it is possible to deduce the type of an individual from the dual process of potential ancestors as follows:

- First, we reconstruct the genealogy of this individual, going backwards in time and looking for its potential ancestors, until we reach the initial condition.
- Then, the individual of the selectively advantageous type if, and only if at least one of his potential ancestors at time 0 is of the advantageous type.
- Otherwise, it is of the selectively disadvantageous type.

In particular, in order to obtain the type of the individual, we do not need to know which one is its true parent among all its potential ancestors.

Conversely, in the model of [DF16], we can distinguish two levels of "genetic diversity":

- An artificial "genetic diversity", which corresponds to types 1/real and 0/ghosts.
- An observable genetic diversity, which corresponds to the potential marks borne by type 1 individuals.

This time, in order to know the complete type of an individual (i.e. including a potential mark), it is not enough to know the types of all its potential ancestors at time 0, and it is also necessary to know which one of them is the true one. This problem will also arise in the variants of the spatial Λ-Fleming Viot process and the Wright-Fisher model that I will introduce later.

In order to answer this question, the authors of [DF16] consider a dual based on a graphical representation of the biased voter model [Gri79; Har78]. Here, I will present instead another dual process, based on the *ordered ASG* used among others in [Len+15], and which keeps an ordering of the different potential ancestors. The interest of this dual process is that its construction can be generalized to other models of expanding populations, such as the ones I constructed during my PhD.

To illustrate this approach, we consider the individual in x at time t, and assume that the last reproduction event which affected it involved the individual in y. If the event was given by $P^{n,(x,y)}$, then we know that the parent is y. However, we assume that the event was given by $\widetilde{P}^{n,(x,y)}$. Then, the individual has two "potential parents": y and x. Moreover, to know its type, we must look at the type of the individual in y, and then at the type of the individual in x if the individual in y was of type 0. Therefore, to keep track of this, we *order* the two potential parents.

Then, we consider the next reproduction event (going backwards in time) to affect one of the two potential ancestors. Again, we assume that it is a "selective" event.

- If the event involves the pair (y, z), $z \neq x$, then we update the ordered list of potential ancestors, and obtain: z, y, x.
- If the event involves the pair (x, z), $z \neq y$, then we obtain: y, z, x.
- If the event involves the pair (x, y), the ordered list does not change.
- If the event involves the pair (y, x), we obtain : x, y.

This last case also highlights that the last potential ancestor in the sequence is not necessarily the true ancestor if all the potential ancestors are of type 0.

This argument is then repeated until we reach the initial condition. It can be used to construct a well-defined dual process, as whenever a selective event occurs, the number of potential ancestors increases by at most one. Thus, the jump rate of the process is bounded by the one of a Yule process. Once the initial condition is reached, we obtain an ordered sequence of potential ancestors x_1, ..., x_n that verifies the following property.

"If at least one of the potential ancestors x_1, ..., x_n at time 0 is of type 1, then the individual in x at the present time is also of type 1, and is a descendant of the first ancestor of type 1 in the ordered sequence. Moreover, it bears the same mark as its ancestor at time 0."

Formally, the dual we consider is defined on the state space

$$\Lambda^{exp} := \left\{ \left(A^{(1)}, ..., A^{(m)} \right) : m \in \mathbb{N}\backslash\{0\} \text{ and for all } l \in [\![1, m]\!],\, A^{(l)} \in \mathrm{Seq}_f \left(L^{-1}\mathbb{Z} \times [\![1, M_n]\!] \right) \right\}$$

where $\mathrm{Seq}_f(L^{-1}\mathbb{Z} \times [\![1, M_n]\!])$ is the set of finite *sequences* of elements of $L^{-1}\mathbb{Z} \times [\![1, M_n]\!]$ in which each element appears only once.

To reconstruct the genealogies of the individuals living in x_1, ..., x_m at time t, we start from the initial condition $((x_1), ..., (x_m))$. Whenever a reproduction event occurs, we update the sequence of potential ancestors of each affected individual, and we go back to the initial condition. Note that the same individual can appear simultaneously in the sequence of potential ancestors of several different individuals from the sample. Defined this way, this process checks the following properties.

Proposition 2.2.1. *Let $m \in \mathbb{N}\backslash\{0\}$, and let $x_1, ..., x_m \in L_n^{-1}\mathbb{Z} \times [\![1, M_n]\!]$. Let $(A_t^{(1)}, ..., A_t^{(m)})_{t\geq 0}$ be the dual process with initial condition $((x_1), ..., (x_m))$ introduced earlier.*

- *For all $t \geq 0$,*

$$\mathbb{E}\left[\prod_{l=1}^{m} \xi_t^n(x_l) \right] = \mathbb{E}\left[\prod_{l=1}^{m} \left(1 - \prod_{x \in A_t^{(l)}} (1 - \xi_0^n(x)) \right) \right].$$

- *For all $t \geq 0$, if we set $A_t^{(l)} = (x_{t,1}^{(l)}, ..., x_{t,|A_t^{(l)}|}^{(l)})$ for all $l \in [\![1, m]\!]$, then*

$$\mathbb{E}\left[\eta_t^n(x_1) \right] = \mathbb{E}\left[\sum_{k=1}^{|A_t^{(1)}|} \eta_0^n(x_{t,k}^{(1)}) \prod_{k'=1}^{k-1} \left(1 - \xi_0^n(x_{t,k'}^{(1)}) \right) \right]$$

and

$$\mathbb{E}\left[\prod_{l=1}^{m} \eta_t^n(x_l) \right] = \mathbb{E}\left[\prod_{l=1}^{m} \left(\sum_{k=1}^{|A_t^{(l)}|} \eta_0^n(x_{t,k}^{(l)}) \prod_{k'=1}^{k-1} \left(1 - \xi_0^n(x_{t,k'}^{(l)}) \right) \right) \right].$$

These equations can be interpreted as follows.

- The sample contains only real individuals if, and only if, each of the individuals has at least one real potential ancestor.

- The sample contains only marked type 1 individuals if and only if the real parent of each individual is of type 1 and marked. Here, we sum over all possible indices for the true parent.

Remark 2.2.2. Using "type 0 individuals" to represent empty areas is an idea that can be found in interacting particle systems, such as the contact process. Referring to them as "ghost individuals" is very visual, but these ghost individuals are first and foremost a modeling artifact. In particular, the genetic diversity that can actually be observed is the diversity among type 1 individuals, or real individuals. This has consequences in terms of the distribution of genealogies: indeed, the set of potential ancestors of a type 1 individual is conditioned to contain at least one real individual, and has a different distribution from the one of the set of potential ancestors of an individual of unknown type (and thus potentially ghost).

This approach has yielded first results on the expansion dynamics and the evolution of genetic diversity at the front of an expanding population. In order to present them, we now introduce the density of type 1 individuals in the patch $i \in L_n^{-1}\mathbb{Z}$ at time t, denoted $u_t^n(i)$. That is, we set

$$u_t^n(i) := \frac{1}{M_n} \sum_{j=1}^{M_n} \xi_t^n(i,j),$$

and we interpolate between the points of $L_n^{-1}\mathbb{Z}$ in order to obtain a function defined over $\mathbb{R}$.

Let $\mathcal{C}_{[0,1]}(\mathbb{R})$ be the set of continuous functions $f : \mathbb{R} \to [0,1]$, equipped with the topology of uniform convergence over compact sets. The following results corresponds to Theorem 1 from [DF16].

Theorème 5. *Assume that u_0^n converges in $\mathcal{C}_{[0,1]}(\mathbb{R})$ towards $f_0 \in \mathcal{C}_{[0,1]}(\mathbb{R})$, and that*

$$\begin{aligned} r_n M_n L_n^{-2} &\xrightarrow[n\to+\infty]{} \alpha \in (0,\infty), \\ M_n R_n^{-1} &\xrightarrow[n\to+\infty]{} \beta \in [0,\infty), \\ r_n L_n^{-1} &\xrightarrow[n\to+\infty]{} \gamma \in [0,\infty), \\ L_n &\xrightarrow[n\to+\infty]{} +\infty, \\ \text{et } L_n R_n &\xrightarrow[n\to+\infty]{} +\infty. \end{aligned}$$

Then, the density in type 1 individuals $(u_t^n)_{t\geq 0}$ converges in distribution in $D([0,\infty), \mathcal{C}_{[0,1]}(\mathbb{R}))$ towards a $\mathcal{C}_{[0,1]}(\mathbb{R})$-valued continuous process $(u_t)_{t\geq 0}$ which is the weak solution to the SDE

$$\begin{cases} \partial_t u &= \alpha \Delta u + 2\theta\beta u(1-u) + \sqrt{4\gamma u(1-u)}\dot{W}, \\ u_0 &= f_0, \end{cases}$$

where $\dot{W}$ is a space-time white noise over $[0,\infty) \times \mathbb{R}$.

Therefore, in the large population limit, under a diffusive rescaling and assuming that the selective advantage of real individuals over ghost individuals is small (i.e., that real individuals slowly invade the space), we recover the Fisher-KPP equation. This strongly suggests that the approach will not be directly applicable to the case of populations living in larger dimensions, and motivates the construction and study of other models for expanding populations.

Theorem 4 from [DF16] also describes the coupled dynamics of marked type 1 individuals in the expanding population. The proofs of the two theorems are based on a proof technique from [MS95], initially applied to a voter model with long range interactions.

The result obtained in [DF16] has since been extended to the case of populations living on a graph in [Fan21].

2.2.2 The ∞-parent SLFV

Integrating ghost individuals to the SLFV

Informally, the SLFV can be seen as modeling the density $\omega_t : \mathbb{R}^d \to [0,1]$ in individuals of one specific type. Integrating ghost individuals to the SLFV can be done by considering that occupied areas contain type 1 individuals, while empty areas contain "ghost" individuals, or type 0 individuals. We obtain a specific case of the SLFV with selection introduced in [FP17], which I will refer as the *k-parent SLFV*.

Informal definition 2.2.3. *(k-parent SLFV) Let $k \in \mathbb{N}\backslash\{0,1\}$, let μ a be σ-finite measure on $(0,+\infty)$ such that*

$$\int_0^\infty \mathcal{R}^d \mu(d\mathcal{R}) < +\infty \tag{2.2.1}$$

and let Π be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,+\infty)$ with intensity $dt \otimes dx \otimes \mu(dr)$. Let $\omega^0 : \mathbb{R}^d \to \{0,1\}$ be a measurable function. Informally, the k-parent spatial Λ-Fleming Viot process $(\omega_{k,t})_{t\geq 0}$ with initial condition ω^0 (or k-parent SLFV) can be defined as follows. For all $(t,x,\mathcal{R}) \in \Pi$, given $\omega_{k,t-}$, k potential parents are sampled uniformly at random in $\mathcal{B}(x,\mathcal{R})$:

- *If at least one of them is of type 1, then for all $y \in \mathcal{B}(x,\mathcal{R})$, we set $\omega_{k,t}(y) = 0$, and we consider that the actual parent is the first real potential parent chosen.*

  This occurs with probability

$$1 - \mathrm{Vol}(\mathcal{B}(x,\mathcal{R}))^{-k} \left(\int_{\mathcal{B}(x,\mathcal{R})} \omega_{k,t-}(y) dy \right)^k .$$

- *Otherwise, for all $y \in \mathcal{B}(x,\mathcal{R})$, we set $\omega_{k,t}(y) = 1$, and the true parent is the last (type 0) potential parent chosen.*

  This occurs with probability

$$\mathrm{Vol}(\mathcal{B}(x,\mathcal{R}))^{-k} \left(\int_{\mathcal{B}(x,\mathcal{R})} \omega_{k,t-}(y) dy \right)^k .$$

See Figure 2.4 for an illustration of the model dynamics.

This definition of the k-parent SLFV is not rigorous, because an infinite number of reproduction events occur at each time step in $\mathbb{R}^d$. Formally, the k-parent SLFV is a measure-valued process, taking its value in the set $\mathcal{M}_\lambda$ of all measures M over $\mathbb{R}^d \times \{0,1\}$ such that there exists $\omega : \mathbb{R}^d \to \{0,1\}$ (and not $[0,1]$) measurable which satisfies

$$M(dx, d\kappa) = ((\omega(x)\delta_0(d\kappa) + (1-\omega(x))\delta_1(d\kappa))dx.$$

A rigorous definition of the k-parent SLFV can be found in Chapter 3.

The dual process associated to the k-parent SLFV, which I will refer to as the *k-parent ancestral process*, corresponds to reconstructing the genealogies of all the potential ancestors of a sample of individuals. Indeed, determining who is the actual parent out of the k potential ones requires the knowledge of the types of *all* the potential ancestors.

Definition 2.2.4. *(k-parent ancestral process) Let $\Xi^0 be \in \mathcal{M}_p(\mathbb{R}^d)$. The k-parent ancestral process $(\Xi_{k,t})_{t\geq 0}$ with initial condition Ξ^0 is the $\mathcal{M}_p(\mathbb{R}^d)$-valued Markov jump process defined as follows. First, let $\Xi_{k,0} = \Xi^0$. Then, let $\overleftarrow{\Pi}$ be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,+\infty)$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R})$. For all $(t,x,\mathcal{R}) \in \overleftarrow{\Pi}$ affecting at least one atom of $\Xi_{k,t-}$, and given $\Xi_{k,t-} = \sum_{i=1}^{N_{t-}} \delta_{x_i}$,*

- *Let $y_1, ..., y_k$ be sampled independently and uniformly at random in $\mathcal{B}(x, \mathcal{R})$.*
- *Let $\mathcal{S}^{\mathcal{R}}(\Xi_{k,t-})$ be the set of all integers $i \in [\![1, N_{t-}]\!]$ such that $x_i \in \mathcal{B}(x, \mathcal{R})$.*
- *Then, we set*

$$\Xi_{k,t} = \Xi_{k,t-} - \sum_{i \in \mathcal{S}^{\mathcal{R}}(\Xi_{k,t-})} \delta_{x_i} + \sum_{i=1}^{k} \delta_{y_i}.$$

This process is well-defined, as its jump rate can be bounded by the one of a Yule process with k descendants.

The duality relation can then be written as follows.

Proposition 2.2.5. *Let $l \in \mathbb{N}^*$, and let ψ be a density over $(\mathbb{R}^d)^l$. If*

$$(\Xi_{k,t})_{t \geq 0} = \left(\sum_{i=1}^{N_{k,t}} \delta_{\xi^i_{k,t}} \right)_{t \geq 0}$$

is a k-parent ancestral process, then for all $t \geq 0$,

$$\mathbb{E}_{\omega_{k,0}=\omega^0} \left[\int_{(\mathbb{R}^d)^l} \psi(x_1, ..., x_l) \left\{ \prod_{j=1}^{l} \omega_{k,t}(x_j) \right\} dx_1 ... dx_l \right]$$
$$= \int_{(\mathbb{R}^d)^l} \psi(x_1, ..., x_l) \mathbb{E}_{\Xi_{k,0}=\Xi[x_1,...,x_l]} \left[\prod_{j=1}^{N_{k,t}} \omega^0(\xi^j_{k,t}) \right] dx_1 ... dx_l.$$

This relation can be interpreted as follows. The probability that l individuals in x_1, ..., x_l at time t are all of type 0 (i.e., that the corresponding spatial positions are empty) is equal to the probability that *all the potential ancestors* of each individual are all of type 0. The proof of the duality relation as well as the construction of the k-parent SLFV are direct adaptations of the proofs in [EVY20], which deals with the case $k = 2$.

Limit $k \to +\infty$ of the k-parent SLFV

The interpretation of type 0 individuals as corresponding to ghost individuals and empty areas encourages us to explore the limit of the k-parent SLFV when $k \to +\infty$. This parameter regime corresponds to giving a very strong selective advantage to real individuals over ghost individuals. The dynamics of the limiting process can be described as follows. Whenever a reproduction event occurs:

- If the affected area contains a non-zero fraction of type 1 individuals, then the actual parent is chosen uniformly at random among all type 1 individuals, and its offspring completely fill the area.
- Otherwise, the area remains entirely empty.

From a modeling viewpoint, the interest of this process comes from the fact that it is akin to the Eden growth model, but continuous in space, and associated to a dual process encoding the genealogies. This motivates the study of the growth properties of the ∞-parent SLFV, in order to compare them to the ones of the Eden model, as we do in the next section and in Chapter 4.

From a mathematical viewpoint, the main characteristic of the ∞-parent SLFV process compared to other SLFVs is directly related to its dual process. Indeed, in the ∞-parent SLFV, whenever a

reproduction event occurs, we need to know the composition of the *entire* affected area in order to be able to decide who is the actual parent. Therefore, instead of being defined on the space $\mathcal{M}_p(\mathbb{R}^d)$ of finite point measures, the dual of the ∞-parent SLFV is this time defined on what can be interpreted as the space of geometric shapes on $\mathbb{R}^d$: the set $\mathcal{M}^{cf}$ of measures of the form

$$\mathcal{M}^{cf} := \{m(E) = \mathbb{1}_{\{x\in E\}}dx : E \in \mathcal{E}^{cf}\},$$

where $\mathcal{E}^{cf}$ is the set of all finite unions of subsets of $\mathbb{R}^d$ which are connected, closed, Lebesgue-measurable and with a finite but nonzero Lebesgue measure. In the case of the dual of the ∞-parent SLFV, these subsets are usually balls.

Let μ be a σ-finite measure on $(0,\infty)$ satisfying Condition (2.2.1). The dual of the ∞-parent SLFV, called the *∞-parent ancestral process*, is defined as follows.

Definition 2.2.6. *(∞-parent ancestral process) Let $E^0 \in \mathcal{E}^{cf}$. The ∞-parent ancestral process $(m(E_t))_{t\geq 0}$ with initial condition $m(E^0)$ is the $\mathcal{M}^{cf}$-valued Markov jump process defined as follows. First, we set $m(E_0) = m(E^0)$. Then, let $\overleftarrow{\Pi}$ be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,+\infty)$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R})$. For all $(t,x,\mathcal{R}) \in \overleftarrow{\Pi}$, if $\mathrm{Vol}(\mathcal{B}(x,\mathcal{R}) \cap E_{t-}) \neq 0$, then*

$$m(E_t) = m(E_{t-} \cup \mathcal{B}(x,\mathcal{R})).$$

In order to show that this process is well-defined, it is no longer possible to compare its jump rate to the one of a Yule process, as we did earlier for the k-parent ancestral process. Indeed, when k is finite, we can use the fact that each atom is affected by a reproduction event at an identical rate. Here, the atoms would correspond to the balls associated to each reproduction event, but the rate at which they are affected by a reproduction event depends on their radius.

When the radius of reproduction events is bounded by $\mathcal{R}_0 > 0$ (i.e., if $\mu((\mathcal{R}_0,\infty)) = 0$), it is still possible to adapt the proof by considering the maximal rate at which a ball is affected by a reproduction event. This allows to bound the jump rate of the process by the one of a Yule process with 2 descendants and conclude.

Conversely, if the radius of reproduction events is not bounded, the process is not necessarily well-defined. Under a technical assumption on μ (introduced in Chapter 3), it is possible to compare the jump rate of the process to the one of a branching process, whose initial condition depends on both E^0 and μ. The non-explosion of the branching process implies the non-explosion of the ∞-parent ancestral process. But the converse is not necessarily true, and we cannot conclude to whether the process is well-defined when μ does not satisfy this condition.

In order to construct the ∞-parent SLFV rigorously, it is possible to use the following proof structure:

1. Define an operator encoding the conjectured dynamics, and construct a solution to the martingale problem associated to this operator.
2. Construct what we conjecture to be the associated dual process.
3. Establish a duality relation between any solution to the martingale problem and the candidate dual process. This allows us to conclude that the solution to the martingale problem is unique.

The duality relation linking the ∞-parent SLFV to its dual has the following form, quite different from the one corresponding to other SLFVs.

Proposition 2.2.7. *Let ω^0 be a measurable function such that $\omega^0 : \mathbb{R}^d \to \{0,1\}$, and let $E^0 \in \mathcal{E}^{cf}$. Let $(\omega_{\infty,t})_{t\geq 0}$ be the ∞-parent SLFV with initial condition ω^0, and let $(m(E_t))_{t\geq 0}$ be the ∞-parent dual process with initial condition E^0. Then, for all $t \geq 0$,*

$$\mathbb{E}\left[\delta_0\left(\int_{E^0}(1-\omega_{\infty,t}(x))\,dx\right)\right] = \mathbb{E}\left[\delta_0\left(\int_{E_t}\left(1-\omega^0(x)\right)dx\right)\right],$$

where $\delta_0 : x \in \mathbb{R} \to \{0,1\}$ is the function equal to 1 if $x = 0$, and equal to 0 otherwise.

In other words, the area E^0 is entirely empty at time t if, and only if the potential ancestors at time 0 of the individuals living in E^0 at time t are (almost) all type 0 individuals.

Notice that because of the second step in the proof, the construction of the ∞-parent SLFV as the unique solution to a martingale problem is only valid if μ satisfies the technical assumption mentioned above, which is stricter than the one ensuring that other SLFVs can be characterized as the unique solutions to martingale problems. However, it is also possible to define the ∞-parent SLFV in a different way, using a sequence of coupled k-parent SLFVs (see Chapter 3, Section 2). This alternative construction has two interests:

- It remains valid even when μ does not verify the technical assumption.
- It explicitly makes the ∞-parent SLFV appear as the limit of the k-parent SLFV when $k \to +\infty$.

The coupling between k-parent SLFVs is based on the construction introduced in [VW15], which "enriches" the information regarding reproduction events given by the Poisson point process. Therefore, we consider an *extended Poisson point process*, in which the locations of the potential parents are sampled along with reproduction events. The approach I used is the following: to each reproduction event, add a sequence $(\mathcal{P}_n)_{n\geq 1} \in \mathcal{B}(0,1)^{\mathbb{N}}$ of positions in the reproduction ball , each position corresponding to a potential parent and being distributed uniformly at random over $\mathcal{B}(0,1)$. If the law followed by $(\mathcal{P}_n)$ is denoted $\tilde{u}$, then we can equivalently consider the Poisson point process defined on $\mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{B}(0,1)^{\mathbb{N}}$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R}) \otimes \tilde{u}(d(p_n))_{n\geq 1}$. This Poisson point process can be used to construct a sequence of k-parent SLFVs, $k \geq 2$ all having the same initial condition, using each time the k *first potential parents* sampled to construct the k-parent SLFV. The coupling ensures that any occupied area in the k-parent SLFV at time t is also occupied in the k'-parent SLFV, $k' \geq k$. Formally, this amounts to saying that

$$\forall t \geq 0, \forall x \in \mathbb{R}^d, \forall k, k' \in \mathbb{N}, k' \geq k \Longrightarrow \omega_{k',t}(x) \leq \omega_{k,t}(x).$$

Therefore, for all $t \geq 0$ and $x \in \mathbb{R}^d$, the sequence $(\omega_{k,t}(x))_{k\geq 2}$ is decreasing. Since it is $\{0,1\}$-valued, it converges towards $\omega_{infty,t}(x) \in \{0,1\}$. The ∞-parent SLFV is then defined as follows.

(Semi) informal definition 2.2.8. *(∞-parent SLFV) The ∞-parent SLFV $(M_t)_{t\geq 0}$ is the unique $\mathcal{M}_\lambda$-valued Markov jump process such that for all $t \geq 0$, M_t satisfies*

$$M(dx, d\kappa) = ((\omega_{\infty,t}(x)\delta_0(d\kappa) + (1 - \omega_{\infty,t}(x))\delta_1(d\kappa))dx.$$

Speed of growth and comparison to the Eden growth model

In this section, we focus on the case of a population living in $\mathbb{R}^2$.

From a modeling viewpoint, the interest of the ∞-parent SLFV comes from the fact that it seems to be a space continuous equivalent of the Eden growth model, but associated to tools allowing one to study genetic diversity. However, this is only a conjecture, and the properties of expansions generated by a ∞-parent SLFV may in fact be very different from those generated by an Eden model.

Therefore, in Chapter 4, we study the growth properties of the region occupied by real individuals in the ∞-parent SLFV, in order to compare them to the ones of the Eden model. In order to do so, we first recall what is known of the growth properties of the Eden model, and more generally of growth properties of first-passage percolation models. The Cox-Durrett shape theorem [CD81] implies that the expansion of a population whose reproduction dynamics follows an Eden model is linear in time. It is possible to obtain upper and lower bounds on the growth rate of the Eden model (see

for example [AP02; BK93]), and more generally on "short-range" percolation models (*short-range percolation*) if the passage time of an edge of the underlying graph is almost surely nonzero (see for instance [ADH17]). However, there is no explicit formula for the expansion speed of Eden's model, and it is necessary to use simulations to obtain an approximation of the expansion speed [AD15]. This is not specific to the Eden growth model, and is in fact also the case for almost all percolation models, with the notable exception of the *corner growth model* (see [Sep09]).

Moreover, the Eden model is conjectured to belong to the equivalence class of the KPZ equation. Again, there is no rigorous proof of this result, neither for the Eden growth model model, nor for almost all probabilistic models of population expansions (with the notable exception of the Solid-on-Solid growth model, see [BG97]).

These reminders about the Eden model have two interests: they illustrate which fundamental properties must be verified by the ∞-parent SLFV in order for us to consider that it is a space continuous equivalent of the Eden growth model, and they give an idea of what kind of results we can expect to prove. In Chapter 4, I show that the expansion of the area occupied by type 1 individuals in the ∞-parent SLFV is linear in time, similarly as in the Eden model. More precisely, I show the following result when there exists $\mathcal{R}_0 > 0$ such that $\mu((\mathcal{R}_0, +\infty)) = 0$, or in other words, when the radius of reproduction events is bounded.

Theorème 6. *Assume that the sub-population of real individuals initially fills the half-plane*

$$\overline{HP}^0 := \{(x, y) \in \mathbb{R}^2 : x < 0\}.$$

Let $(M_t^{HP})_{t\geq 0}$ *be the* ∞*-parent SLFV with initial condition*

$$M_0^{HP}(dz) := \mathbb{1}_{\{z\in\overline{HP}^0\}}dz.$$

Let $(\omega_t^{HP})_{t\geq 0}$ *be a density of* $(M_t^{HP})_{t\geq 0}$*. For all* $x \geq 0$*, we set*

$$\overrightarrow{\tau}_x := \min\left\{t \geq 0 : \lim_{\epsilon\to 0} V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz < 1\right\},$$

where $\mathcal{B}_\epsilon((x, 0))$ *is the ball with center* $(x, 0)$ *and radius* ϵ*, and where* V_ϵ *is the volume of* $\mathcal{B}_\epsilon((x, 0))$*. Then, there exists* $\nu > 0$ *such that*

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overrightarrow{\tau}_x]}{x} = \nu.$$

Intuitively, $\overrightarrow{\tau}_x$ can be interpreted as the time at which the point $(x, 0)$ is reached by the sub-population of real individuals. This theorem means that when the radius of reproduction events is bounded, the expansion of the sub-population of real individuals along a line is linear in time. The corresponding speed is equal to ν^{-1}, which has no explicit expression. This result remains true if we replace balls by *ellipses*, and could easily be generalized to other geometric shapes.

The proof of this result relies on the use of the dual process associated to the ∞-parent SLFV, the ancestral ∞-parent process. Indeed, a location is occupied by type 1 individuals if and only if a nonzero fraction of its potential ancestors at time 0 belong to the half-plane $\overline{HP}^0$. Using the invariance by translation and rotation of the Poisson point process encoding reproduction events and associated to the dual process, the previous theorem is then equivalent to the following result.

Proposition 2.2.9. *For all* $x > 0$*, let* HP^x *be the half-plane:*

$$HP^x := \{(x', y) \in \mathbb{R}^2 : x' \geq x\}.$$

Moreover, let $(E_t)_{t\geq 0}$ *be the* ∞*-parent ancestral process with initial condition* $\{(0,0)\}$ *(which requires to modify the definition of the ancestral process in order to allows for points as initial conditions), and let*

$$\overleftarrow{\tau}_x := \min\{t \geq 0 : \mathrm{Vol}(E_t \cap HP^x) > 0\}.$$

Then, there exists $\nu > 0$ *such that*

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} = \nu.$$

In order to prove this result, we first show that the expansion is *at least* linear in time, that is, that there exists $\nu \geq 0$ (instead of $\nu > 0$) such that

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} = \nu.$$

In order to do so, we consider the times at which each of the half-planes $(HP^{4\mathcal{R}_0 n})_{n\in\mathbb{N}}$ are reached for the first time, and for all $n \in \mathbb{N}$, we set

$$T_{0,n} = \overleftarrow{\tau}_{4n\mathcal{R}_0}.$$

As the radius of reproduction events is bounded by $\mathcal{R}_0$, it is (almost surely) not possible to reach simultaneously the half-planes $HP^{4n\mathcal{R}_0}$ and $HP^{4n'\mathcal{R}_0}$, $n \neq n'$ for the first time.

Then, for all $m \in \mathbb{N}$, when the half-plane $HP^{4m\mathcal{R}_0}$ is reached for the first time, we choose a point of the form $(4m\mathcal{R}_0, y)$, $y \in \mathbb{R}$ in the set $E_{T_{0,m}} \cap HP^{4m\mathcal{R}_0}$, and we restart a ∞-parent ancestral process from this location. This ancestral process is constructed using the same underlying Poisson point process, but only considering reproduction events occuring after time $T_{0,m}$. This is made possible by the fact that the distribution of the Poisson point process is invariant by translation in time.

Let $T_{m,n}$, $n \geq m$ be the time needed by this new ancestral process to reach the half-plane $HP^{4n\mathcal{R}_0}$. Then,

$$\forall n \in \mathbb{N}, n \geq m \Longrightarrow T_{0,n} \leq T_{0,m} + T_{m,n}.$$

The family $(T_{m,n})_{0\leq m\leq n}$ satisfies all the hypothesis of Theorem 1.10 from [Lig85], which gives the existence of $\nu' \geq 0$ such that

$$\lim_{n\to+\infty} \frac{\mathbb{E}[T_{0,n}]}{n} = \nu'.$$

We conclude using the fact that $(\overleftarrow{\tau}_x)_{x\geq 0}$ is increasing.

In the second part of the proof, we show that the expansion is *at most* linear in time, by comparing the dual process to a first-passage percolation problem. The comparison relies on a partitioning of the space into *cells* with side length $2\mathcal{R}_0$, placed on a grid and distant of $6\mathcal{R}_0$. We use these cells to discretize the ∞*-parent ancestral process*, by considering that a cell is *active* if it intersects the ancestral process, and is *inactive* otherwise. The ancestral process starting from a point can be seen as a union of balls with radius bounded by $\mathcal{R}_0$, and not all of them intersect a cell, but they are all necessarily at a finite distance from an active cell. Thus, the growth rate of the ∞-parent ancestral process is the same as the one of the discretized process.

It is this discretized process that we use to construct the comparison with the first-passage percolation problem. Indeed, observe that a cell can only become active if one of the eight nearest neighboring cells is also active. Moreover, when one of the neighboring cells becomes active for the first time, the focal cell can only (potentially) become active after having being affected by a reproduction event, which occurs after a time exponentially distributed. Therefore, we can compare the active and inactive cells in the discretized process to the empty and occupied sites in a "short-range" first-passage percolation problem in which each edge is crossed after a time exponentially distributed. We conclude by using the fact that expansions in such percolation problems are linear in time [CD16].

In order to complete the theoretical study, we also simulated the ∞-parent SLFV and obtained an approximation of the growth rate of the area occupied by real individuals. The simulations showed that the growth rate was much higher than initially conjectured, due to the emergence of "spikes" pointing in the direction of the expansion. These spikes are relatively rare, but when they occur, they then "thicken" in the direction transverse to the expansion, leading to a sudden advance of the front over large distances. See Figure 2.5 for an illustration of how spikes make the front advance faster. In order to investigate this question from a theoretical viewpoint, The last section of Chapter 4 considers a toy model for which it is possible to obtain an explicit expression for the expansion speed, using the invariant distribution of a time-discretized version of the process.

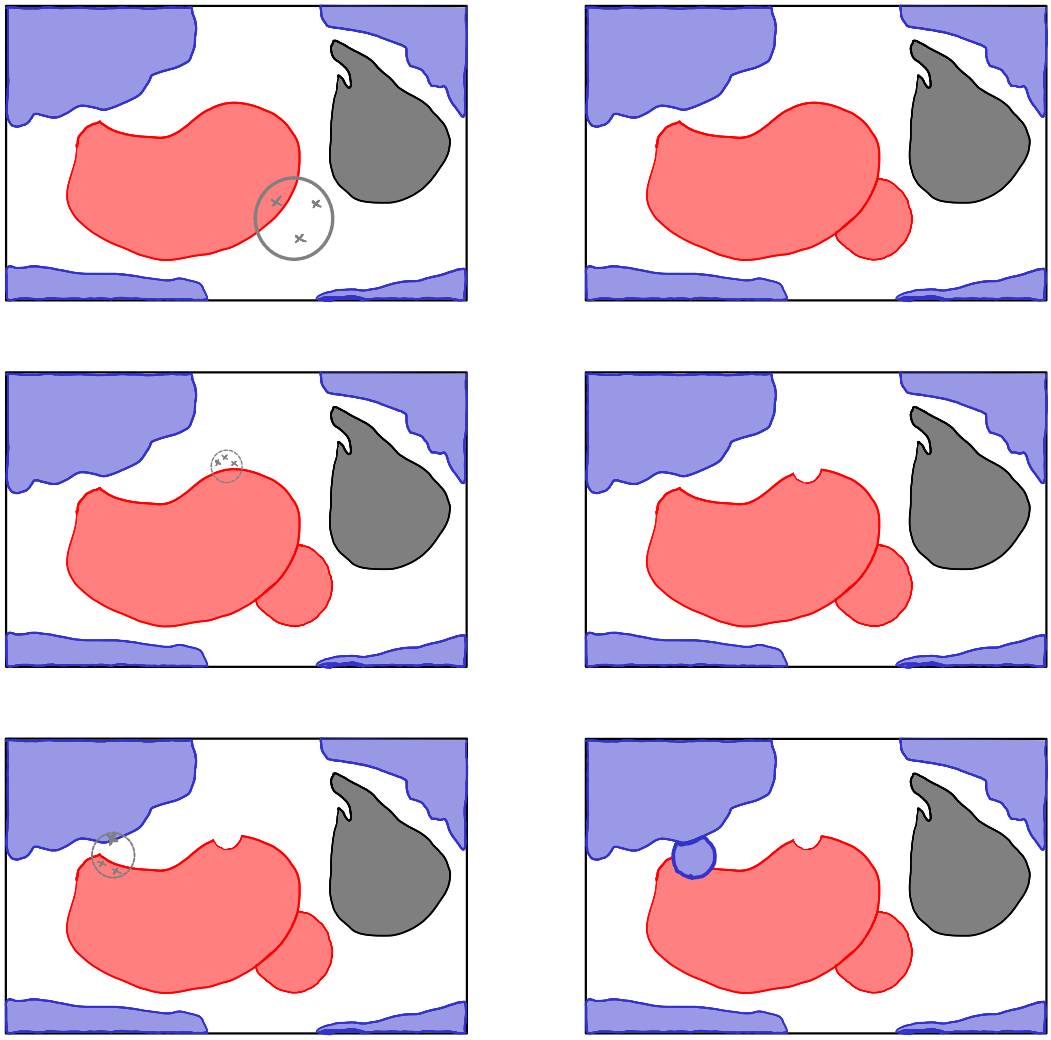


Figure 2.4: Informal definition of the k-parent SLFV, here when $k = 3$. We assume that initially, individuals at the same location are of the same type. However, in the case of the k-parent SLFV, white areas are empty, and modeled as occupied by ghost individuals. (a), (c), (e) We wait for a reproduction event to affect the area. Then, we sample 3 potential parents. (b) If only one of the 3 potential parents is real, then it is the true parent. It reproduces, and its descendants fill the ball. (d) If all 3 potential parents are ghosts, then the ball is filled with ghost individuals. This corresponds to a local extinction event. (f) If at least two potential parents are real, then the actual parent is the first real potential parent chosen. Here we assume that the actual parent is the blue parent.

Figure 2.5: Growth of the occupied area in the ∞-parent SLFV, starting from an initial condition in which the real individuals initially fill the half-plane just above the image. The images represent the state of the same sub-population after increasingly longer expansion times. Each color represents a different sub-type among real individuals. The white areas correspond to empty areas, modeled as occupied by ghost individuals.

2.2.3 An extension of the Wright-Fisher model for plant metapopulations living in a fragmented and disturbed environment

We now consider how to integrate the concept of ghost individuals to the Wright-Fisher model. As a reminder, the goal is construct an extension of the Wright-Fisher model with a spatial structure which would be adapted to the study of plant dynamics in urban tree bases. Therefore, the model must include the following elements:

- frequent and localized extinction events,
- a seed bank.

In order to do so, we first introduce the spatially structured Wright-Fisher model that we will use as a basis to construct the complete model. We consider that the metapopulation is composed of an infinite number of patches, indexed by $i \in \mathbb{Z}$ and located along a line. Each patch can contain $M \in \mathbb{N}\backslash\{0\}$ plants. In order make it easier to integrate a seed bank component later, we assume that at the beginning of each generation, the M plants are present as seeds. These M seeds all germinate, and grow into plants, which will in turn produce new seeds in the following way: each seed from patch i chooses a parent uniformly at random among the plants in the patch i with probability $1-2c$, $c \in (0, 1/2)$, or among the plants in the patch $i-1$ (resp. $i+1$) with probability c.

We now add ghost individuals to the model. This means that seeds and plants can be of two types : type 1, corresponding to real individuals, and type 0, or "ghost", encoding empty areas in a patch. The reproduction dynamics is modified as follows. Instead of choosing a single parent, each seed in patch i chooses $k \in \mathbb{N}\backslash\{0,1\}$ *potential parents*, each taken from patch i with probability $1-2c$ and from patch $i-1$ (resp. $i+1$) with probability c. The seed is then of type 1 if at least one of its potential parents is of type 1 (in this case, the actual parent is the first potential parent of type 1 chosen), and of type 0 otherwise. As before, the parameter k quantifies the "selective advantage" of real individuals over ghost individuals, and hence the speed at which real individuals colonize a new environment.

The addition of ghost individuals makes it easy to model extinction events as synchronized "mutations" of all individuals in a patch to type 0. We assume that extinction events occur after the germination phase, but before the seed production phase. When a patch is affected by an extinction event, all the plants it contains die and become ghost plants, i.e. type 0 plants. At each generation, each patch is affected by an extinction event independently and with probability $p \in [0, 1]$.

In order to integrate the seed bank component, we combine ideas from [Bla+16] and [KKL01]. Indeed, we are interested in plants, that is, in species whose seeds can generally only remain dormant without losing viability over rather short durations, of the order of ten years at most (though there are exceptions). This is closer to the assumptions underlying the model from [KKL01], but the model introduced in [Bla+16] is easier to study due to being Markovian. Using ghost individuals allows us to introduce the idea of a loss of viability of seeds after having been dormant for too long into the model of [Bla+16].

As before, we assume that each patch contains exactly M seeds of type 0 or 1, or in other words, at most M real seeds. Each of these seeds can remain dormant during at most $H \in \mathbb{N}$ complete generations without losing viability. If it remains dormant longer, then it "mutates" and becomes a ghost (or type 0) seed.

Then, at each generation, only $\lfloor gM \rfloor$, $g \in (0, 1)$ seeds germinate. These seeds are chosen uniformly at random from the seed bank, regardless of their type and the time already spent in the seed bank. After the potential extinction events, $\lfloor gM \rfloor$ new seeds are produced and integrate the seed bank, while the plants die.

We have now built a model for plant metapopulations in urban tree bases, which is more generally suited to plant metapopulations living in a fragmented and disturbed environment. I will refer

to this model as the k-parent WFSB (for "*Wright-Fisher metapopulation process with a Seed Bank component*"). See Figure 2.6 for an illustrated summary of the model dynamics.

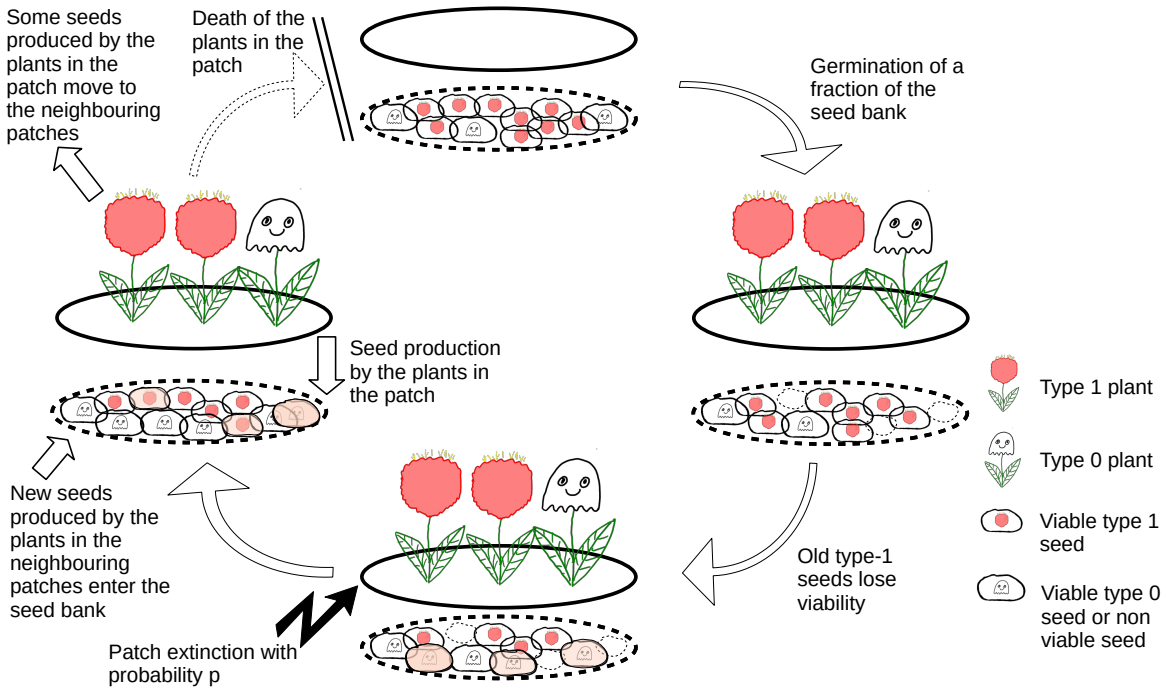


Figure 2.6: Illustration of the intra-patch dynamics of the k-parent WFSB. Here, $\lfloor gM \rfloor = 3$ et $M = 12$.

Formally, the model is defined on the state space $\mathcal{F}_M \times \mathcal{H}_M$, where

$$\mathcal{F}_M := \left\{ (\xi_{i,j})_{i\in\mathbb{Z},j\in[\![1,M]\!]} \in \{0,1\}^{\mathbb{Z}\times[\![1,M]\!]} : \mathrm{Card}\left(\{(i,j) \in \mathbb{Z}\times[\![1,M]\!] : \xi_{i,j} = 1\}\right) < +\infty \right\},$$

and $\mathcal{H}_M := \left\{ (h_{i,j})_{i\in\mathbb{Z},j\in[\![1,M]\!]} : \forall (i,j) \in \mathbb{Z}\times[\![1,M]\!], h_{i,j} \in \mathbb{N} \right\}$.

For all $(i,j) \in \mathbb{Z}\times[\![1,M]\!]$, $h_{i,j}$ represents the number of complete generations that the seed occupying the seed bank compartment j of patch i already spent dormant, and $\xi_{i,j}$ represents its type *when it was produced*. In other words, the current type of this seed is given by $\xi_{i,j}\mathbb{1}_{\{h_{i,j}\leq H\}}$. The model is then defined as follows.

Definition 2.2.10. *(k-parent WFSB metapopulation process) Let $(\xi,h) \in \mathcal{F}_M \times \mathcal{H}_M$. The k-parent Wright-Fisher metapopulation process with seed bank, with parameters (M,H,g,c,p) and initial condition (ξ,h) and denoted by $(\xi^n,h^n)_{n\in\mathbb{N}}$, is the $(\mathcal{F}_M\times\mathcal{H}_M)$-valued Markov chain defined by $(\xi^0,h^0) = (\xi,h)$ and for all $n \in \mathbb{N}$, given (ξ^n,h^n) :*

1. *For each $i \in \mathbb{Z}$, we sample $\lfloor gM \rfloor$ different seed bank compartments $s_{i,1},...,s_{i,\lfloor gM\rfloor} \in [\![1,M]\!]$ uniformly at random in patch i.*
2. *Let $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ be i.i.d $\{0,1\}$-valued random variables such that $\mathbb{P}(\mathrm{Ext}_1 = 1) = p$.*
3. *For all $(i,j) \in \mathbb{Z}\times[\![1,M]\!]$, if $j \in \{s_{i,j'} : j' \in [\![1,\lfloor gM\rfloor]\!]\}$, we first set $h_{i,j}^{n+1} = 0$. Moreover, let $C_{1,i,j},...,C_{k,i,j}$ be i.i.d $\{-1,0,1\}$-valued random variables such that*

$$\mathbb{P}(C_{1,i,j} = 1) = \mathbb{P}(C_{1,i,j} = -1) = c,$$

and such that $(C_{l,i,j})_{1\leq l\leq k}$ *is independent from* $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ *and from* $(C_{i',j',l})_{1\leq l\leq k}$ *for* $(i',j')\neq(i,j)$*.*

For all $l\in[\![1,k]\!]$*, if* $\mathrm{Ext}_{i+C_{l,i,j}}=1$*, we set* $\tilde{k}_l=0$*, and if* $\mathrm{Ext}_{i+C_{l,i,j}}=0$*, we sample one seed bank compartment* j_l *uniformly at random among the* $\lfloor gM\rfloor$ *ones sampled in the patch* $i+C_{l,i,j}$ *(those in the set* $\{s_{i+C_{l,i,j},j'}:j'\in[\![1,\lfloor gM\rfloor]\!]\}$*), and we set*

$$\tilde{k}_l=\xi^n_{i+C_{l,i,j},j_l}\mathbb{1}_{\{h^n_{i+C_{l,i,j},j_l}\leq H\}}.$$

We conclude by setting $\xi^{n+1}_{i,j}=\max\{\tilde{k}_l:l\in[\![1,k]\!]\}$*.*

4. *On the other hand, if* $j\notin\{s_{i,j'}:j'\in[\![1,\lfloor gM\rfloor]\!]\}$*, we set* $\xi^{n+1}_{i,j}=\xi^n_{i,j}$ *and* $h^{n+1}_{i,j}=h^n_{i,j}+1$*.*

It is possible to define a dual process associated to the k-parent WFSB, again following the potential ancestors of each individual in a sample. This time, however, we do not necessarily need to reconstruct the complete genealogy of a potential parent to know its type. Indeed, if the plant chosen as a potential parent is in a patch that was just affected by an extinction event, then we know that it is a type 0 plant. Similarly, if the plant comes from a seed produced more than $H+1$ generations ago, then the seed was non-viable (hence of type 0) when it germinated, and the corresponding plant is of type 0. Therefore, the dual process is akin to the Ancestral Selection Graph (or ASG), but to which some branches are *pruned* as the result of extinction events or seeds losing viability. This construction is based on the *pruned ASG* introduced in [Len+15]. I will not define this process rigorously, however, as the results in Chapter 6 do not rely on the use of the dual process.

During my PhD, I have focused on the study of the *forwards-in-time* process, in order to understand under which conditions an expansion is possible. In order to do so, I made use of the following observation. If real plants are present in patch i at generation n, then these plants can produce seeds that will potentially integrate the seed banks of patches $i-1$, i and $i+1$, and germinate during generations $n+1$, ..., $n+H+1$. We can thus define a notion of *reachable patches*. Moreover, some of these patches will in fact be affected by extinction events, and will not be able to contain real plants. In other words, if we consider the space $\mathbb{Z}\times\mathbb{N}$, where $(i,n)\in\mathbb{Z}\times\mathbb{N}$ corresponds to patch i during generation n :

1. Each site $(i,n)\in\mathbb{Z}\times\mathbb{N}$ is *extinct* with probability p, independently from other types.
2. Starting from site $(i,n)\in\mathbb{Z}\times\mathbb{N}$, it is possible to reach sites $(i+i',n+n')$, $i'\in\{-1,0,1\}$, $n'\in[\![1,H+1]\!]$.

We obtain an oriented site-percolation problem, which has already been studied in the litterature. Using a result from [HS21], we can deduce the existence of a critical patch extinction probability $p_{crit}(H)\in(0,1)$, depending on the parameter H and above which population expansions are not possible, no matter the values taken by the other parameters.

Theorem 2.2.11. *For all* $H\in\mathbb{N}$*, there exists* $p_{crit}(H)\in(0,1)$ *such that for all* $M\in\mathbb{N}\backslash\{0\}$*,* $k\in\mathbb{N}\backslash\{0,1\}$*,* $g\in(0,1)$ *and* $c\in(0,1/2)$*, for all* $(\xi,h)\in\mathcal{F}_M\times\mathcal{H}_M$ *and* $p\in(p_{crit}(H),1]$*, if* $(\xi^n,h^n)_{n\in\mathbb{N}}$ *is the* k*-parent WFSB with initial condition* (ξ,h) *and with parameters* (M,H,g,c,p)*, then*

$$\lim_{n\to+\infty}\mathbb{P}\left(\forall(i,j)\in\mathbb{Z}\times[\![1,M]\!],\xi^n_{i,j}\mathbb{1}_{\{h^n_{i,j}\leq H\}}=0\right)=1.$$

The proof relies on the use of a process called the BOA process (or "Best Occupancy Achievable" process), which encodes which patches can potentially contain real seeds given the initial

condition and which patches were affected by extinction events. The BOA process can be interpreted as an oriented site-percolation problem, and is defined over the state space $\mathcal{F}^\infty \times \mathcal{H}^\infty$, where

$$\mathcal{F}^\infty := \{(O_i)_{i\in\mathbb{Z}} : \forall i \in \mathbb{Z}, O_i \in \{0,1\} \text{ and } \mathrm{Card}\left(\{i \in \mathbb{Z} : O_i = 1\}\right) < +\infty\}$$
$$\text{and } \mathcal{H}^\infty := \{(h_i)_{i\in\mathbb{Z}} : \forall i \in \mathbb{Z}, h_i \in \mathbb{N}\}.$$

Définition 2.2.12. *(BOA process) Let $(O,h) \in \mathcal{F}^\infty \times \mathcal{H}^\infty$. The BOA process $(O^{\infty,n}, h^{\infty,n})_{n\in\mathbb{N}}$ with initial condition (O,h) is the $\mathcal{F}^\infty \times \mathcal{H}^\infty$-valued Markov chain such that $(O^{\infty,0}, h^{\infty,0}) = (O,h)$ and for all $n \in \mathbb{N}$, given $(O^{\infty,n}, h^{\infty,n})$:*

1. *Let $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ be i.i.d. random variable, each following a Bernouilli distribution with probability of success p.*

2. *For all $i \in \mathbb{Z}$, if $\mathrm{Ext}_i = 0$ and $O_i^{\infty,n}\mathbb{1}_{\{h_i^{\infty,n}\leq H\}} = 1$, then we set*

$$O_{i-1}^{\infty,n+1} = O_i^{\infty,n+1} = O_{i+1}^{\infty,n+1} = 1$$
$$\text{and} \qquad h_{i-1}^{\infty,n+1} = h_i^{\infty,n+1} = h_{i+1}^{\infty,n+1} = 0.$$

   We do not do anything during this step otherwise.

3. *For all $i \in \mathbb{Z}$, if $O_i^{\infty,n+1}$ was not already defined during step 2, we set $O_i^{\infty,n+1} = O_i^{\infty,n}$ and $h_i^{\infty,n+1} = h_i^{\infty,n} + 1$.*

If the BOA process is constructed using the "same" initial condition and the same extinction events as the k-parent WFSB, then for all $n \in \mathbb{N}$ and $i \in \mathbb{Z}$:

- If $O_i^{\infty,n} = 0$, then patch i cannot contain viable type 1 seeds at the beginning of generation n.
- If $O_i^{\infty,n} = 1$, this means that if patch i contains type 1 seeds at the beginning of generation n, then they are at least of age $h_i^{\infty,n}$. In particular, if $h_i^{\infty,n} > H$, then the patch cannot contain viable type 1 seeds.

Therefore, it is possible to construct a coupling between the k-parent WFSB and the BOA process in such a way so that the k-parent WFSB is "included" in the BOA process. Under this coupling, the extinction of the BOA process implies the extinction of the k-parent WFSB. We can then use results from percolation theory to deduce the existence of a patch extinction probability $p_{crit}(H)$ above which the k-parent WFSB almost surely goes extinct.

The inclusion of the k-parent WFSB in the BOA process is generally strict, and deviations may occur in one of the following three cases :

1. Type 1 plants are indeed present in a patch, but none of them are chosen as potential parents.
2. Type 1 seeds are present in a patch, but none of them germinate.
3. Type 1 seeds entered the seed bank less than $H+1$ generations ago, but all of them already germinated.

However, when $M \to +\infty$ and $k \to +\infty$, if k grows "faster" than M (i.e., if there exists $\alpha > 1$ such that $k = \lceil M^\alpha \rceil$), then the probability of each of these events tends to 0, and we can show that the k-parent WFSB converges to the BOA process.

Theorem 2.2.13. *Let $\alpha > 1$. For all $M \geq 2$, let $(\xi^{(M),n}, h^{(M),n})_{n\in\mathbb{N}}$ be the $\lceil M^\alpha \rceil$-parent WFSB with parameters (M, H, g, c, p), and let $(O^{(M),\infty,n}, h^{(M),\infty,n})_{n\in\mathbb{N}}$ be the associated BOA process. For all $M \geq 2$, $i \in \mathbb{Z}$ and $n \in \mathbb{N}$, we set*

$$O_i^{(M),n} := 1 - \prod_{j\in[\![1,M]\!]} \left(1 - \xi_{i,j}^{(M),n}\right)$$

$$h_i^{(M),0} := \begin{cases} \min\{h_{i,j}^{(M),0} : j \in [\![1, M]\!] \text{ and } \xi_{i,j}^{(M),0} = 1\} & \text{if } O_i^{(M),0} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{et } h_i^{(M),n} := \begin{cases} \min\{h_{i,j}^{(M),n} : j \in [\![1, M]\!] \text{ and } \xi_{i,j}^{(M),n} = 1\} & \text{if } O_i^{(M),n} = 1 \\ h_i^{(M),n-1} + 1 & \text{otherwise.} \end{cases}$$

Then, for all $N \in \mathbb{N}$,

$$\mathbb{P}\left(\bigcap_{n=0}^{N} \left(\left\{\forall i \in \mathbb{Z}, O_i^{(M),n} = O_i^{(M),\infty,n}\right\} \cap \left\{\forall i \in \mathbb{Z}, h_i^{(M),n} = h_i^{(M),\infty,n}\right\}\right)\right) \xrightarrow[M\to+\infty]{} 1.$$

This convergence result allows us to conclude that $p_{crit}(H)$ can indeed be considered as the critical patch extinction probability for the k-parent WFSB.

2.3 SPOM metapopulation models and seed banks

2.3.1 What is a SPOM?

The convergence of the k-parent WFSB to the BOA process is also of interest from a metapopulation modeling viewpoint. Indeed, it bridges the gap between two families of metapopulation models: individual-based models, whose biological interpretation is more direct but whose mathematical analysis is generally difficult, and *Stochastic Patch Occupancy Models*, or SPOMs. These Markovian models are characterized by the fact that they only encode the presence or absence of the focal species in each patch, as does the BOA process. Therefore, the BOA process can be interpreted both as a model encoding the effect of extinction events on the species distribution, or as a simplified metapopulation dynamics model.

Assuming that metapopulation dynamics depend only on the species presence/absence represents a strong simplification of intra-patch dynamics, which has already been discussed in the literature (see for example [Bag04; DSV03; Han04]). This approximation is considered suitable for species living in highly fragmented environments [Han04] such as tree bases in urban areas [DPC11; Oma+19]. Moreover, it makes data collection easier (it is indeed easier to collect presence/absence data than abundance data), and allows to obtain theoretical results on these models. See for example [HO03] for a presentation of several of these results. Moreover, since SPOMs are Markov chains, estimators could be developed for several of the classical SPOMs [Moi99; Moi04].

In order to check the validity of this assumption, several studies [AN94; Kee02; OH04] have already focused on comparing the dynamics of a complete individual-based model to the one of a SPOM approximation, through the use of simulations. They concluded that under the parameter regimes studied, the SPOM approximation generally approaches well the dynamics observed in the individual-based model. The convergence result of the k-parent WFSB to the BOA process completes these different studies, by giving a theoretical result rather than one based on simulations.

We now briefly introduce some classical SPOMs models. In order to do so, we first recall the central idea behind metapopulation theory: in a fragmented environment, local extinction events (i.e., at the scale of a single patch) are frequent, due to external disturbances or to the small patch

population size. However, recolonization events are also frequent, and hence the species does not go extinct on a regional scale. The *Propagule Rain Model*, or PRM [Got91], and the Levins model [Lev69] are two classical SPOMs models that can be considered as two extremes regarding how colonization is modeled. Indeed, in the PRM, new individuals come from an external source (the propagule rain giving its name to the model), while in the Levins model, they can only come from other patches.

The original version of the Levins model can be seen as an island model, in the sense that the geographical distance between patches does not play a part in colonization dynamics. As a result, new individuals can come from a nearby patch as well as a more distant patch, with the same probability. This model has since been modified to include a spatial structure, and make colonization probabilities depend on the distance between patches (see for instance [Moi04]).

Formally, the PRM and the Levins model with a special structure (or *spatially realistic Levins model*) are defined as follows. Assume that patch locations correspond to the discrete (usually finite) set $E \subset \mathbb{R}^d$, $d \geq 1$ (the most common case being $d = 2$), and for any $(i, j) \in E^2$, let $d_{i,j}$ denote the distance between patches i and j. The PRM and the Levins model are each defined on the space $\{0, 1\}^E$, and are characterized by the following parameters:

- PRM MODEL: *colonization probability* c, *intrinsic extinction probability* p.
- SPATIALLY REALISTIC LEVINS MODEL: *mean dispersal distance* $\delta \in (0, \infty)$, *intrinsic extinction probability* p, *connectivity parameter* $\gamma \in (0, \infty)$.

These two models are defined as follows.

Définition 2.3.1. *(Propagule Rain Model) Let $(x_i^0)_{i\in E} \in \{0, 1\}^E$. The Propagule Rain Model $(X^n)_{n\in\mathbb{N}}$ with initial condition $X^0 = (x_i^0)_{i\in E}$ and parameters (c, p) is the $\{0, 1\}^E$-valued Markov chain whose probability transitions are defined as follows. For all $n \in \mathbb{N}$, given $X^n = (X_{i'}^n)_{i'\in E}$, for all $i \in E$,*

- *If $X_i^n = 0$, then $X_i^{n+1} = 1$ with probability c, and $X_i^{n+1} = 0$ otherwise.*
- *If $X_i^n = 1$, then $X_i^{n+1} = 0$ with probability $p(1 - c)$, and $X_i^{n+1} = 1$ otherwise.*

In other words, in the PRM, during each generation, if a patch is extinct, it becomes occupied with probability c. If the patch is occupied, it becomes empty with probability p, but is then immediately recolonized with probability c. Therefore, the extinction is only visible with probability $p(1 - c)$.

Définition 2.3.2. *(Spatially realistic Levins model) Let $(x_i^0)_{i\in E} \in \{0, 1\}^E$. The spatially realistic Levin model $(X^n)_{n\in\mathbb{N}}$ with initial condition $X^0 = (x_i^0)_{i\in E}$ and with parameters (δ, γ, p) is the $\{0, 1\}^E$-valued Markov chain whose probability transitions are defined as follows. For all $n \in \mathbb{N}$, given $X^n = (X_{i'}^n)_{i'\in E}$, for all $i \in E$,*

- *If $X_i^n = 0$, then $X_i^{n+1} = 1$ with probability*

$$C_i^{n+1}(\delta, \gamma) = 1 - \exp\left(-\gamma \sum_{j\in E\setminus\{i\}} X_j^n \exp\left(-\delta^{-1} d_{i,j}\right)\right),$$

  and $X_i^{n+1} = 0$ otherwise.
- *If $X_i^n = 1$, then $X_i^{n+1} = 0$ with probability $p(1 - C_i^{n+1}(\delta, \gamma))$, and $X_i^{n+1} = 1$ otherwise.*

Contrary to the PRM, in the (spatially realistic) Levins model, new seeds come from other patches, and colonization probabilities depends on the distance between patches. In particular, the higher the number of neighbouring occupied patches, the higher the colonization probability.

2.3.2 Accounting for seed bank presence and statistical inference

While SPOMs have been successfully applied to the study of insect metapopulations (see for example [Han11; MSH98]) or small mammals (see for example [Ozg+06]), their application to plant metapopulations has been hampered by the fact that classical SPOM models do not take into account the potential presence of a seed bank in the soil [FW02]. Estimating model parameters while neglecting the potential presence of dormant seeds introduces an important bias on colonization and extinction rates [Fré+13], and increases the rate of (false) identification of a PRM, as a seed bank leaves a imprint close to the one of a propagule rain [Oma+19]. Recently, several SPOMs models taking into account the presence of a seed bank have been developed [Bor+15; Fré+13; Plu+18]. These models encode the presence/absence of plants as well as viable seeds. They are all Hidden Markov models (or HMMs): the presence/absence of seeds is a *hidden state*, which is not visible but influences the presence/absence of plants in the patch (the *observed state*). Therefore, it can be inferred from observations. The BOA process can also be seen as a HMM. Indeed, the age of the youngest viable seeds in the seed bank of a patch is a hidden state, which affects whether plants can be observed in the patch.

The model introduced in [Plu+18] is a seed bank variant of the PRM. It is associated to an estimator based on a variant of the EM algorithm [DLR77] for HMMs. It has been successfully applied to real presence/absence data of weeds in agroecosystems [Plu+18] and in vineyards [Kaz+21]. Assuming that patch locations are again given by $E \subset \mathbb{R}^d$, this variant of PRM is this time defined on the state space $\{(0,0),(1,0),(1,1)\}^E$. If $(z_i)_{i\in E} = (s_i, x_i)_{i\in I} \in \{(0,0),(1,0),(1,1)\}^E$ corresponds to the state of the metapopulation during generation $n \in \mathbb{N}$, then for all $i \in E$, $s_i = 1$ if the seed bank of patch i contains (viable) seeds at the beginning of generation n, and $x_i = 1$ if patch i contains plants during generation n. In the PRM with a seed bank component, colonization occurs after germination, so it is not possible to have $s_i = 0$ and $x_i = 1$.

The model is characterized by three parameters:

- the colonization probability c;
- the probability of germination and survival of seedlings to adulthood, denoted g;
- the seed bank death probability d.

Formally, it is defined as follows.

Definition 2.3.3. *(PRM with a seed bank component) Let $(z_i^0)_{i\in E} \in \{(0,0),(1,0),(1,1)\}^E$. The PRM with a seed bank component $(Z^n)_{n\in\mathbb{N}}$ with initial condition $Z^0 = (z_i^0)_{i\in E}$ and parameters (c,g,d) is the $\{(0,0),(1,0),(1,1)\}^E$-valued Markov chain whose transition probabilities are as follows. For all $n \in \mathbb{N}$ and $i \in E$, given $Z^n = (Z_i^n)_{i\in E}$,*

- *If $Z_{i'}^n = (0,0)$, then*
$$Z_{i'}^{n+1} = \begin{cases} (0,0) & \text{with probability } 1-c, \\ (1,0) & \text{with probability } c(1-g), \\ (1,1) & \text{with probability } cg. \end{cases}$$
- *If $Z_{i'}^n = (1,0)$, then*
$$Z_{i'}^{n+1} = \begin{cases} (0,0) & \text{with probability } d(1-c), \\ (1,0) & \text{with probability } (1-d(1-c))(1-g), \\ (1,1) & \text{with probability } (1-d(1-c))g. \end{cases}$$
- *If $Z_{i'}^n = (1,1)$, then*
$$Z_{i'}^{n+1} = \begin{cases} (1,0) & \text{with probability } 1-g, \\ (1,1) & \text{with probability } g. \end{cases}$$

This definition can be interpreted as follows. If a patch does not contain any seeds, then it is colonized with probability c, and the corresponding seeds germinate with probability g. If the patch does contain seeds, either at least some of them germinate (with probability g), and grow into plants that will produce new seeds, or none of them germinate. In the second case, the seed bank may then die with probability d, which is immediately compensated by a recolonization with probability c. See Figure 2.7 for an illustration of the model dynamics.

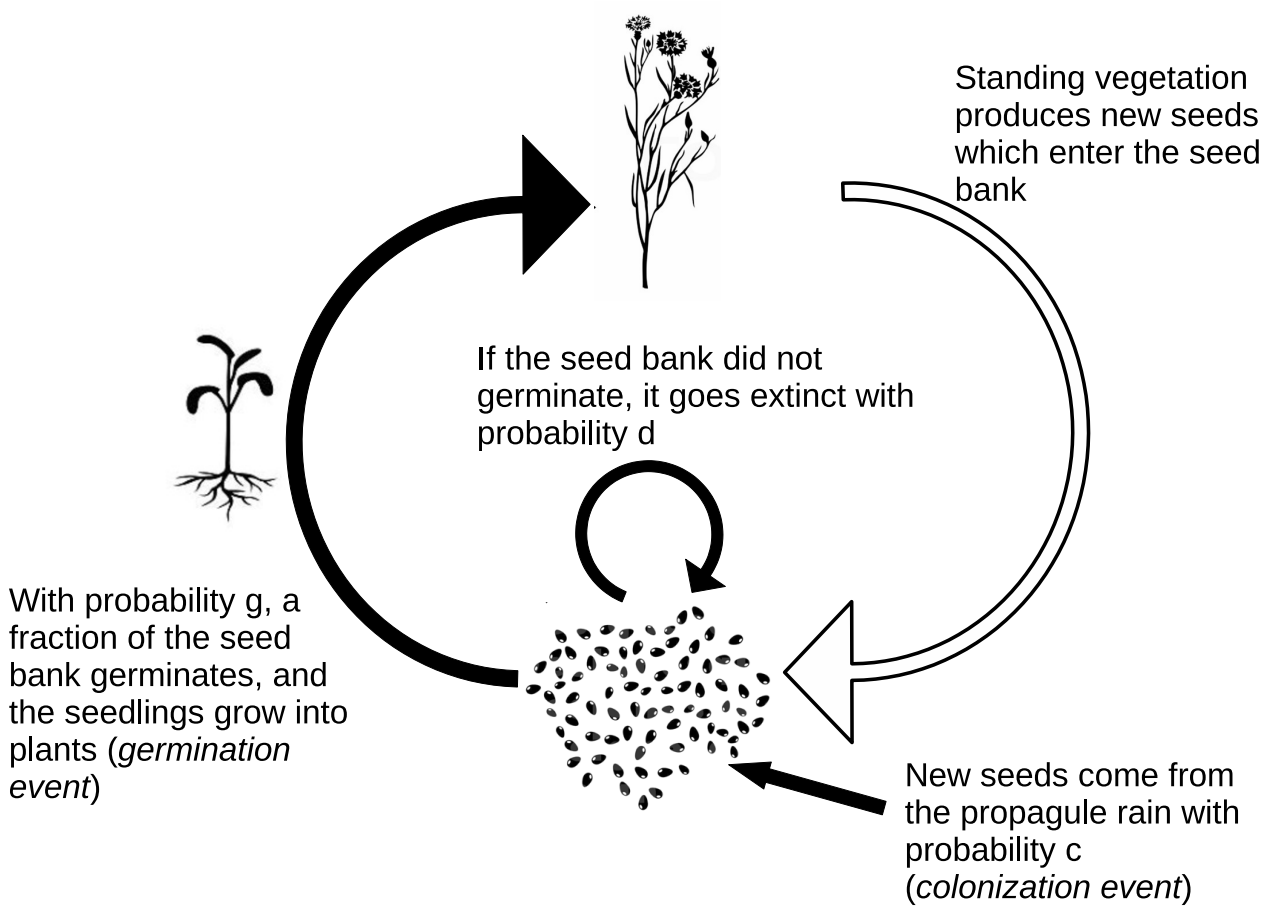


Figure 2.7: Illustration of the dynamics of the PRM with seed bank.

During my PhD, one of my goals was to apply the associated estimator to the dataset of plants in urban tree bases in the 12th administrative district of Paris, described above (see Section 1.1.2), in order to confirm the results of [Oma+19] regarding the presence of a seed bank in urban tree bases. However, it was not possible to use directly the estimator introduced in [Plu+18] as the amount of data available was not sufficient to be able to differentiate the model with a seed bank from the model without a seed bank, even though the dataset comprises 10 years of surveys, each streets ranging from 30 to 200 tree bases. Therefore, this problem is likely to occur with other smaller datasets.

The introduction of a new metric, called the *Seed Bank Characteristic Event probability*, or SBCE probability, allowed us to circunvent this problem. This metric, presented in detail in Chapter 5, measures the contribution of the potential seed bank to the observed dynamics. It is equal to zero when there is no seed bank (i.e. when $d = 1$), but also when the seed bank has no visible effect on the observed dynamics (i.e. when $c = 1$ and/or $g = 1$). Moreover, when a seed bank is present, the SBCE probability is lower when colonization events are frequent and extinction events are rare, that is, when the presence of the seed bank has little influence on the observed dynamics, and therefore

when it is more difficult to conclude to the presence/absence of a seed bank from plant observations. Concretely, the SBCE probability corresponds to the probability that the plants present in a patch produce seeds that will go dormant before germinating, without a colonization or seed bank death event occurring inbetween. Thus,

$$\mathbb{P}_{SBCE} = g\frac{(1-g)(1-c)(1-d)}{1-(1-g)(1-c)(1-d)}.$$

The main contribution of Chapter 5 is to show that although this metric is less informative than knowing whether a seed bank is actually present, it requires in practice less data to be sufficiently accurate for the applications considered. Moreover, it is less sensitive to survey errors, making it applicable to datasets from citizen science, for instance. Depending on the ecological issue of interest, the SBCE probability can therefore represent an interesting metric.

From a theoretical viewpoint, the approach highlights the relevance of focusing on the detectability of the seed bank (in particular in SPOM-type models), and of taking this aspect into account in order to assess the performance of a estimator. Moreover, the application of the SBCE probability to the dataset of plants in urban tree bases uncovered a significant contribution of the seed bank to the observed dynamics for some plant species. The methodology used can also be applied to other datasets.

2.4 Future prospects

During my PhD, I developed two families of population genetics models for expanding populations:

- a family of models continuous in space and time, based on the spatial Λ-Fleming Viot process,
- a family of models discrete in space and time, adapted to the modeling of plant metapopulations in a fragmented environment affected by frequent local extinction events.

My PhD works focus on the construction of these models, and on the development of tools for the reconstruction of genealogies of samples of individuals. The logical continuation of this work would be to use these models to study genetic diversity at the front of expanding populations. This would require the development of tools allowing one to make use of the dual processes encoding genealogies.

In order to do so, I plan to explore the following two avenues. First of all, I will study the genealogies of individuals *conditioned to be real*, i.e. observable individuals. Indeed, if an individual is real, then at least one of its potential ancestors in the initial condition is real too. Moreover, there is an ancestral lineage composed only of real individuals, which goes back to the actual ancestor of the individual in the initial condition. In the absence of mutations, this ancestor carries the same marked type as the individual under consideration.

Rather than searching all the potential ancestors of a real individual, we can focus instead on this specific ancestral line, which is conditioned to go back to the area from which the expansion started. Moreover, in the case of the k-parent WFSB, it is conditioned to avoid extinction events and to remain dormant during at most H consecutive generations. After having studied the distribution of the "true" ancestral lineage of a real individual, the next step would be to investigate how multiple lineages associated to different individuals interact. The results obtained could be used to simulate these lineages, or to build estimators to make inference from real data.

Then, another possible approach is to study the model through simulations. This approach faces a major obstacle: the intuitive way to code the process is not very efficient from a computational point of view. This is especially true for the k-parent SLFV and the ∞-parent SLFV, due to the

construction based on a Poisson point process. Simulating the process more efficiently would require looking at alternative constructions or representations of the process, or exploiting theoretical results on the model dynamics. Being able to simulate the ∞-parent SLFV more efficiently could for instance allow to refine the conjecture that the model belongs to the equivalence class of the KPZ equation, in addition to making it possible to study the emergence of sectors at the front of expanding populations.

Although my PhD work is mostly related to population genetics models, it also contains contributions to the modeling and study of metapopulations. In particular, one of the limiting processes obtained, the BOA process, can be interpreted as a SPOM-type model. It can be seen as a simplified version of the spatially structured Levins model, also including the potential presence of a seed bank. Therefore, it completes the variety of already existing SPOM-type models. A work in progress, in collaboration with Nathalie Machon and Amandine Véber, focuses on the development of an estimator associated with the BOA process allowing statistical inference from plant presence/absence data. The estimation method relies on classical techniques for Hidden Markov models, adapted to the particular case of the BOA process. We will use the estimator in association with the one associated with the Propagule Rain Model (or PRM) with a seed bank component, and apply it to the dataset of plants in urban tree bases. The goal is to identify where seeds enter the system, and which areas correspond to ecological corridors. The methodology used could be applied more generally to metapopulations of plants in which the presence of a seed bank is suspected.

2.5 Outline

My thesis comprises two parts. The first one, composed of Chapters 3 and 4, focuses on population genetics models for expanding populations based on the spatial Λ-Fleming Viot process, or SLFV. Chapter 3 deals with the integration of the concept of "ghost individuals" to the spatial Λ-Fleming Viot process. The resulting process is called k-parent SLFV, in which the parameter k corresponds to the number of potential parents chosen. Chapter 3 then focuses on the construction of the limit of this process when $k \to +\infty$, yielding the ∞-parent SLFV. This chapter is based on a paper submitted to *ESAIM - Probability and Statistics*. Chapter 4 focuses on the growth properties of the ∞-parent SLFV, and shows that the expansion of the area occupied by real individuals is linear in time. In order to understand how the distinctive dynamics observed at the expansion front can lead to a faster expansion than initially conjectured, it includes the study of a toy model for which it is possible to obtain an explicit expression for the growth rate. This chapter is based on a joint work with Amandine Véber, which was submitted to *Electronic Journal of Probability*.

The second part of my thesis is more applied and focuses on how seed banks influence metapopulation dynamics of plants living in a fragmented environment characterized by frequent local extinction events, in particular in the case of plants in urban tree bases. Chapter 5 is a preliminary study conducted on the dataset of floristic inventories of plants in urban tree bases described earlier. It is a joint work with Nathalie Machon, Jean-Baptiste Mihoub and Alexandre Robert, which has been published in *Methods in Ecology and Evolution*. Chapter 6 deals with the construction and study of a corresponding theoretical model, based on a variant with ghost individuals of the Wright-Fisher model with a seed bank component. The main result of this chapter is the existence of a critical patch extinction probability depending on seed bank parameters above which a population expansion is not possible. The corresponding paper has been accepted for publication in *Theoretical Population Biology*.

Part II

A probabilistic population genetics model for expanding populations in continuous space

Chapter 3

The k-parent spatial Lambda-Fleming-Viot process as a stochastic measure-valued model for an expanding population

This chapter is based on the article [Lou21], available on Arxiv and submitted to ESAIM: Probability and Statistics.

Abstract

We model spatially expanding populations by means of a spatial Λ-Fleming Viot process (SLFV) with selection : the k-parent SLFV. We fill empty areas with type 0 "ghost" individuals, which have a strong selective disadvantage against "real" type 1 individuals. This model is a special case of the SLFV with selection introduced in [EVY20; FP17] : natural selection acts during all reproduction events, and the fraction of individuals replaced during a reproduction event is constant equal to 1. Letting the selective advantage k of type 1 individuals over type 0 individuals grow to $+\infty$, and without rescaling time nor space, we obtain a new model for expanding populations, the ∞-parent SLFV. This model is reminiscent of the Eden growth model [Ede61], but with an associated dual process of potential ancestors, making it possible to investigate the genetic diversity in a population sample. In order to obtain the limit $k \to +\infty$ of the k-parent SLFV, we introduce an alternative construction of the k-parent SLFV adapted from [VW15], which allows us to couple SLFVs with different selection strengths.

3.1 Introduction

Population expansions are common events occuring at all biological scales. The growth of a population in a new environment generates interfaces with distinctive features [Hue+10; KPZ86] and specific patterns of genetic variation [Gra+13; Hal+07; HN10], both being a consequence of the stochasticity of reproduction at the front, where local population sizes are small. The models which are used to study expanding populations can be divided in two main categories : growth models, mostly used to investigate the front features, and models coming from population genetics, which are more suited to study genetic diversity patterns.

Experimental approaches suggest that the dynamics of fronts of real expanding populations belongs to the universality class of the Kardar-Parisi-Zhang (KPZ) equation introduced in [KPZ86] (see e.g [Hue+10]). It has been conjectured (and demonstrated in the case of the solid-on-solid (SOS) growth model [BG97]) that many growth models generate similar interfaces. One of these models is the Eden growth model, initially introduced on a lattice in [Ede61]. Under this model, each node of the lattice is either occupied or empty. At each time step, an empty node with at least one occupied neighbour is chosen, and becomes occupied. There exist alternative update rules for this model [JB87], as well as off-lattice variants (see e.g [WLB95]). While this model can be used to study the growth of an expanding population, it is less suited to study genetic diversity patterns.

Conversely, models used in population genetics are generally associated with tools allowing one to investigate these patterns. The analysis of the genetic diversity of a population often goes through modelling the ancestral lineages of a subset of individuals, and studying how these lineages coalesce into common ancestors [Eth11]. However, most classical population genetics models assume that populations have constant sizes and that individuals are uniformly distributed over the area of interest. Therefore, they appear at first ill-suited to model a population during an expansion event. One way to overcome this consists in filling empty areas with "ghost" individuals, which can reproduce but have a very strong selective disadvantage against "real" individuals [DF16; HN08]. Under this framework, the reproduction of "ghost" individuals can be interpreted as a local extinction of real individuals.

Using this idea, it is possible to model a population expansion as the spread of a genetic type favoured by natural selection. Such a question was already studied by means of different models including a stochastic component, mostly in one dimension (see e.g [Bar+13; EP20; Kor+10]). The most classical one is based on the Fisher-KPP equation [Fis37; KPP37], in which stochasticity is introduced through a Wright-Fisher noise term. If $0 \leq p(t,x) \leq 1$ represents the proportion of individuals of the favoured type at location $x \in \mathbb{R}$ at time $t \geq 0$, then $p(t,x)$ solves the stochastic Fisher-KPP equation if for $x \in \mathbb{R}$ and $t > 0$,

$$\frac{\partial p}{\partial t}(t,x) = \frac{m}{2}\Delta p(t,x)dt + s_0 p(t,x)(1-p(t,x)) + \sqrt{\frac{1}{p_e}p(t,x)(1-p(t,x))}W(dt,dx) \tag{3.1.1}$$

where W is a space-time white noise and p_e an effective population density. In one dimension, the stochastic Fisher-KPP equation exhibits travelling wave solutions [MS95], which describe how does the advantageous type spreads through space. However, Eq. (3.1.1) has no solution in higher dimensions. Many variants of the deterministic version of the Fisher-KPP equation have been studied, including versions with individuals having different motilities [Bou+12; Cal+18], different growth rates (see e.g [Def+19]), other diffusion kernels, or other choices for the nonlinearity [BNss].

Other models are more individual-based, and are adaptations of classical population genetics models, such as the Moran model [DF16; EP20; HN08] or the stepping-stone model [Aus+97; Bar+13; PE15]. They generally require to divide the space into subunits called *demes*, to which reproduction events are limited and which are connected by migration.

In this article, we will focus instead on a "reproduction event-based" model allowing us to keep the spatial continuum : the spatial Λ-Fleming Viot process, or SLFV [BEV10; Eth08]. Its main feature

is that it models reproduction events affecting whole areas rather than reproduction individual by individual, by means of a Poisson point process of reproduction events.

The original version of the SLFV does not account for the presence of a selectively favoured genetic type, but it can be modified in order to incorporate selection : see [FP17] for different forms of fixed selection mechanisms, and [BEK21; CK19; KR20] for ways to introduce fluctuating selection. Our approach will be based on a version of the SLFV with selection introduced in [FP17] and rigourously constructed in [EVY20]. Most of the work on the SLFV with selection involved investigating scaling limits under different forms of *weak* selection (see also [Eth+17; EFS17; EFP17]). However, in our case, since the selectively disadvantaged individuals do not actually exist, the selection can be considered as very strong. Therefore, we shall consider a different limit, when selection goes stronger and stronger, and neither time nor space are rescaled. The limiting model we shall obtain will be close to the off-lattice Eden growth model, hence we can expect it to generate interfaces similar to the ones observed in real expanding populations. Moreover, and contrary to the off-lattice Eden growth model, a dual process of potential ancestors is also associated to it, giving us tools to investigate the genetic diversity patterns observed in an expanding population. This model therefore constitutes a new model for expanding populations, which naturally appears as the limit of other well-known processes, and which seems promising in order to investigate both the front features and the genetic diversity patterns of an expanding population. In this work, we will focus on the area occupied by the population, but future works will include genetic diversity inside the expanding population, using for instance tracer dynamics [DF16; HN08].

3.1.1 The k-parent SLFV and its dual

The k-parent SLFV

All the random objects introduced in this section will be defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Before presenting the processes we will consider, we need to introduce some notation.

Let $d \geq 1$. Let $C_c(\mathbb{R}^d)$ be the space of all continuous and compactly supported functions $\mathbb{R}^d \to \mathbb{R}$, let $C^1(\mathbb{R})$ be the space of all continuously differentiable functions on $\mathbb{R}$, let $C_b^1(\mathbb{R})$ be the space of all bounded functions $\mathbb{R} \to \mathbb{R}$ that are C^1 and whose first derivative is also bounded, and let $\mathcal{B}(\mathbb{R}^d)$ be the space of all measurable functions $\mathbb{R}^d \to \mathbb{R}$.

We start by introducing the state space over which the variant of the SLFV with selection we consider is defined. Let $\widetilde{\mathcal{M}}_\lambda$ be the space of all measures M on $\mathbb{R}^d \times \{0,1\}$ such that for all $f \in C_c(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d \times \{0,1\}} f(x) M(dx, d\kappa) = \int_{\mathbb{R}^d} f(x) dx.$$

In other words, $\widetilde{M}_\lambda$ is the space of all measures on $\mathbb{R}^d \times \{0,1\}$ whose marginal over $\mathbb{R}^d$ is Lebesgue measure. By a standard decomposition theorem (see e.g [Kal06], p.561), for all $M \in \widetilde{\mathcal{M}}_\lambda$, there exists $\omega : \mathbb{R}^d \to [0,1]$ measurable such that

$$M(dx, d\kappa) = ((\omega(x)\delta_0(d\kappa) + (1 - \omega(x))\delta_1(d\kappa))dx. \tag{3.1.2}$$

Such a ω is not unique, but defined up to a Lebesgue null set. The state space we consider is the set $\mathcal{M}_\lambda$ of all measures $M \in \widetilde{\mathcal{M}}_\lambda$ such that there exists a measurable function $\omega : \mathbb{R}^d \to \{0,1\}$ (instead of $\omega : \mathbb{R}^d \to [0,1]$) satisfying (3.1.2).

We endow $\widetilde{\mathcal{M}}_\lambda$ and $\mathcal{M}_\lambda$ with the topology of vague convergence. Moreover, let $D_{\mathcal{M}_\lambda}[0, +\infty)$ (resp. $D_{\widetilde{\mathcal{M}}_\lambda}[0, +\infty)$) denote the space of all càdlàg $\mathcal{M}_\lambda$-valued paths (resp. $\widetilde{\mathcal{M}}_\lambda$-valued paths), endowed with the standard Skorokhod topology.

Let $M \in \mathcal{M}_\lambda$, and let $\omega : \mathbb{R}^d \to \{0,1\}$ be a measurable function satisfying Eq. (3.1.2). The function ω can be interpreted as the indicator function of a measurable set $E \subset \mathbb{R}^d$ corresponding

to the area occupied by what will be called "type 0" individuals, while $\mathbb{R}^d \backslash E$ corresponds to the area occupied by "type 1" individuals. We will consider that type 0 individuals correspond to the "ghost" individuals mentionned in the introduction, and type 1 individuals to the "real" individuals. Therefore, type 0 individuals have a strong selective disadvantage against type 1 individuals, and E corresponds to the area not yet invaded by the real population (up to a Lebesgue null set). In all that follows, any $\omega : \mathbb{R}^d \to \{0,1\}$ such that (3.1.2) is true will be called a *density* of M, and the notation ω_M will be used to denote an arbitrarily chosen density of M.

For all $f \in C_c(\mathbb{R}^d)$, $F \in C^1(\mathbb{R})$ and $\omega : \mathbb{R}^d \to \{0,1\}$ measurable, we set :

$$\langle \omega, f \rangle := \int_{\mathbb{R}^d} f(x)\omega(x)dx$$

and we define the function $\Psi_{F,f} \in C_b(\mathcal{M}_\lambda)$ as :

$$\begin{aligned} \forall M \in \mathcal{M}_\lambda, \Psi_{F,f}(M) &:= F\left(\int_{\mathbb{R}^d \times \{0,1\}} f(x)\mathbb{1}_{\{\kappa=0\}} M(dx, d\kappa)\right) \\ &= F\left(\int_{\mathbb{R}^d} f(x)\omega_M(x)dx\right) \\ &= F\left(\langle \omega_M, f\rangle\right). \end{aligned} \tag{3.1.3}$$

For all $f \in C_c(\mathbb{R}^d)$, we denote the support of f by $Supp(f)$, and for all $\mathcal{R} \in \mathbb{R}_+^*$, we set :

$$\begin{aligned} Supp^{\mathcal{R}}(f) &:= \{y \in \mathbb{R}^d : \exists x \in Supp(f), ||y - x|| \leq \mathcal{R}\} \\ \text{and} \qquad V_{\mathcal{R}} &:= \mathrm{Vol}(\mathcal{B}(0, \mathcal{R})). \end{aligned}$$

In other words, $V_{\mathcal{R}}$ is the volume of a ball of radius $\mathcal{R}$, and $Supp^{\mathcal{R}}(f)$ is the set of all points which are at a distance of at most $\mathcal{R}$ of a point in the support of f.

For all $\omega : \mathbb{R}^d \to \{0,1\}$, $\mathcal{R} \in \mathbb{R}_+^*$ and $x \in \mathbb{R}^d$, we define the functions $\Theta^+_{x,\mathcal{R}}(\omega) : \mathbb{R}^d \to \{0,1\}$ and $\Theta^-_{x,\mathcal{R}}(\omega) : \mathbb{R}^d \to \{0,1\}$ by :

$$\begin{aligned} \Theta^+_{x,\mathcal{R}}(\omega) &:= \mathbb{1}_{\mathcal{B}(x,\mathcal{R})^c} \times \omega + \mathbb{1}_{\mathcal{B}(x,\mathcal{R})}, \\ \Theta^-_{x,\mathcal{R}}(\omega) &:= \mathbb{1}_{\mathcal{B}(x,\mathcal{R})^c} \times \omega. \end{aligned}$$

$\Theta^+_{x,\mathcal{R}}(\omega)$ corresponds to filling the ball $\mathcal{B}(x,\mathcal{R})$ with type 0 individuals (or equivalently, emptying the ball $\mathcal{B}(x,\mathcal{R})$ of all real individuals), while $\Theta^-_{x,\mathcal{R}}(\omega)$ can be interpreted as filling the ball $\mathcal{B}(x,\mathcal{R})$ with type 1 individuals. Notice that if $M \in \mathcal{M}_\lambda$, then $\Theta^+_{x,\mathcal{R}}(\omega_M) \in \mathcal{M}_\lambda$ and $\Theta^-_{x,\mathcal{R}}(\omega_M) \in \mathcal{M}_\lambda$.

We now introduce the operator which will be used to define the specific SLFV with selection we will consider as the solution to a well-posed martingale problem. Let $k \in \mathbb{N}\backslash\{0,1\}$, and let μ be a σ-finite measure on $\mathbb{R}_+^*$ such that

$$\int_0^\infty \mathcal{R}^d \mu(d\mathcal{R}) < +\infty. \tag{3.1.4}$$

Let $\mathcal{L}^k_\mu$ be the operator acting on functions of the form $\Psi_{F,f}$ with $f \in C_c(\mathbb{R}^d)$ and $F \in C^1(\mathbb{R})$, defined

in the following way. Let $f \in C_c(\mathbb{R}^d)$ and $F \in C^1(\mathbb{R})$. Then, for all $M \in \mathcal{M}_\lambda$,

$$\mathcal{L}_\mu^k \Psi_{F,f}(M) := \int_{\mathbb{R}^d} \int_0^\infty \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V_{\mathcal{R}}^k} \times \Bigg[\left(\prod_{j=1}^k \omega_M(y_j) \right) \times F(\langle \Theta_{x,\mathcal{R}}^+(\omega_M), f \rangle) + \left(1 - \prod_{j=1}^k \omega_M(y_j) \right) \times F(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f \rangle) - F(\langle \omega_M, f \rangle) \Bigg] dy_1 ... dy_k \mu(d\mathcal{R}) dx.$$

In Section 3.5, it is shown that this operator is well-defined, and that it can be rewritten as

$$\mathcal{L}_\mu^k \Psi_{F,f}(M) = \int_0^\infty \int_{Supp^{\mathcal{R}}(f)} \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V_{\mathcal{R}}^k} \times \Big[\prod_{j=1}^k \omega_M(y_j) \times F(\langle \Theta_{x,\mathcal{R}}^+(\omega_M), f \rangle) + (1 - \prod_{j=1}^k \omega_M(y_j)) \times F(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f \rangle) - F(\langle \omega_M, f \rangle) \Big] dy_1 ... dy_k dx \mu(d\mathcal{R}).$$

Intuitively, an interpretation of this operator in terms of reproduction events is the following. Whenever a reproduction event affects the ball $\mathcal{B}(x, \mathcal{R})$, k positions $y_1, ..., y_k$ are sampled inside the ball, and we take k individuals occupying each one of these positions. Since the density of type 0 individuals ω_M is $\{0, 1\}$-valued, we can consider that all the individuals occupying the position y_1 (resp. $y_2, ..., y_k$) are of type $1 - \omega_M(y_1)$ (resp. $1 - \omega_M(y_2)$, ..., $1 - \omega_M(y_k)$). If $\prod_{j=1}^k \omega_M(y_j) = 1$, then all the individuals are of type 0, and we fill the ball $\mathcal{B}(x, \mathcal{R})$ with type 0 individuals. Conversely, if $1 - \prod_{j=1}^k \omega_M(y_j) = 1$, then at least one individual is of type 1, and this time we fill the ball $\mathcal{B}(x, \mathcal{R})$ with type 1 individuals. Since type 0 individuals model "ghost" individuals, they are supposed to have a selective disadvantage against "real" type 1 individuals, hence the exclusion of the case $k = 1$ which would not give any advantage to type 1 individuals. Moreover, k can be interpreted as measuring the strength of the selective advantage of "real" individuals against "ghost" individuals, or in other words, the capacity of "real" individuals to invade an empty environment.

If $k = 2$, $\mathcal{L}_\mu^2$ is the operator introduced in [EVY20] to define and characterize the "selection part" of the SLFV with selection, in the special case for which there are no neutral events and all reproduction events have an impact of $u = 1$. Their proof of the existence and uniqueness of the $D_{\widetilde{\mathcal{M}_\lambda}}[0, +\infty)$-valued solution to the martingale problem associated to $\mathcal{L}_\mu^k$ can easily be extended to the case $k \geq 2$, by restricting the martingale problem to an increasing sequence of compact subsets of $\mathbb{R}^d$ converging to $\mathbb{R}^d$. In Section 3.2, we will show that this unique solution is in fact $D_{\mathcal{M}_\lambda}[0, +\infty)$-valued if the initial value belongs to $\mathcal{M}_\lambda$.

Theorem 3.1.1. *Let $k \geq 2$, and let μ be a σ-finite measure on $(0, +\infty)$ satisfying condition (3.1.4). For all $M^0 \in \mathcal{M}_\lambda$, there exists a unique $D_{\mathcal{M}_\lambda}[0, +\infty)$-valued process $(M_t)_{t \geq 0}$ such that $M_0 = M^0$ and, for all $F \in C^1(\mathbb{R})$ and $f \in C_c(\mathbb{R}^d)$,*

$$\left(\Psi_{F,f}(M_t) - \Psi_{F,f}(M_0) - \int_0^t \mathcal{L}_\mu^k \Psi_{F,f}(M_s) ds \right)_{t \geq 0}$$

is a martingale. Moreover, the process $(M_t)_{t \geq 0}$ is Markovian, and the corresponding semigroup is Feller.

Here, uniqueness is meant as uniqueness in distribution. The proof of Theorem 3.1.1 is a straightforward adaptation of the proof of Theorem 1.2 from [EVY20], combined with Lemma 3.2.6. Indeed, Theorem 1.2 from [EVY20] states that the martingale problem $(\mathcal{L}^2_\mu, \delta_{M^0})$ admits a unique solution with values in $\widetilde{\mathcal{M}}_\lambda$ (rather than $\mathcal{M}_\lambda$), which turns out to be Markovian and with Feller semi-group (we recall that $D_{\widetilde{\mathcal{M}}_\lambda}[0,+\infty)$ and $D_{\mathcal{M}_\lambda}[0,+\infty)$ are both equipped with the standard Skrorokhod topology, while $\widetilde{\mathcal{M}}_\lambda$ are $\mathcal{M}_\lambda$ are equipped with the topology of vague convergence). The proof of the uniqueness of the solution is based on Proposition 1.7 from [EVY20], which can be easily extended to the case $k \geq 2$, as done in Proposition 3.1.8. Note that the notation used in this chapter is slightly inconsistent with the one used in [EVY20]: the set denoted as $\mathcal{M}_\lambda$ in Theorem 1.2 corresponds to the set $\widetilde{\mathcal{M}}_\lambda$ in this chapter. Then, in Lemma 3.2.6, we exhibit a $\mathcal{M}_\lambda$-valued solution to the martingale problem $(\mathcal{L}^k_\mu, \delta_{M^0})$, allowing us to conclude.

Definition 3.1.2 (Definition of the k-parent SLFV)**.** *Let $k \geq 2$, let μ be a σ-finite measure on $(0,\infty)$ satisfying (3.1.4), and let $M^0 \in \mathcal{M}_\lambda$. Then, the k-parent spatial Λ-Fleming-Viot process (or k-parent SLFV) with initial condition M^0 associated to μ is the unique solution to the martingale problem $(\mathcal{L}^k_\mu, M^0)$ stated in Theorem 3.1.1. In particular, the k-parent SLFV is a strong Markov process with càdlàg paths a.s. (up to a modification).*

By extension, if $\omega : \mathbb{R}^d \to \{0,1\}$ is measurable, we will define the k-parent SLFV with initial density ω associated to μ to be the k-parent SLFV with initial condition M^0 associated to μ, with $M^0 \in \mathcal{M}_\lambda$ of density ω.

Intuitively, the k-parent SLFV can be constructed in the following way. Let $M^0 \in \mathcal{M}_\lambda$, and let μ be a σ-finite measure on $(0,\infty)$ satisfying (3.1.4). Moreover, let Π be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0,+\infty)$ with intensity $dt \otimes dx \otimes \mu(dr)$. Initially, the k-parent SLFV is equal to M^0. The dynamics of the k-parent SLFV $(M_t)_{t\geq 0}$ is then as follows. If $(t,x,\mathcal{R}) \in \Pi$, a reproduction event happens at time t in the ball $\mathcal{B}(x,\mathcal{R})$. We sample k types according to the type distribution in the ball $\mathcal{B}(x,\mathcal{R})$ at the time $t-$. We interpret these types as the types of k potential "parents". With probability

$$\frac{1}{V_{\mathcal{R}}^k} \int_{\mathcal{B}(x,\mathcal{R})^k} \left[\prod_{j=1}^k \omega_{M_{t-}}(y_j) \right] dy_1 ... dy_k,$$

the k types sampled are 0, so the k potential parents are of type 0. In this case, all the individuals in the ball $\mathcal{B}(x,\mathcal{R})$ die, the k-th potential parent (of type 0) fills the ball $\mathcal{B}(x,\mathcal{R})$ with its descendants, which means that we set :

$$\forall z \in \mathcal{B}(x,\mathcal{R}), \omega_{M_t}(z) = 1.$$

Conversely, with probability

$$1 - \frac{1}{V_{\mathcal{R}}^k} \int_{\mathcal{B}(x,\mathcal{R})^k} \left[\prod_{j=1}^k \omega_{M_{t-}}(y_j) \right] dy_1 ... dy_k,$$

at least one of the k types sampled is 1. As in the other case, all the individuals in the ball $\mathcal{B}(x,\mathcal{R})$ die, but this time the first potential parent to be of type 1 fills the ball $\mathcal{B}(x,\mathcal{R})$ with its descendants, which amounts to setting

$$\forall z \in \mathcal{B}(x,\mathcal{R}), \omega_{M_t}(z) = 0.$$

Note that the position of the parent which actually reproduces is then uniformly distributed over the region $\{y \in \mathcal{B}(x,\mathcal{R}) : \omega_{M_{t-}}(y) = 0\}$. The value taken by the density out of the ball $\mathcal{B}(x,\mathcal{R})$ at time t is not affected by this reproduction event. We repeat this for each $(t,x,\mathcal{R}) \in \Pi$.

This construction can be made rigourous using arguments adapted from [VW15], and will be used in Section 3.2 to complete the proof of Theorem 3.1.1.

Remark 3.1.3. The k-parent SLFV is a special case of the general definition of an SLFV with selection in [FP17], with impact parameter $u = 1$, selection parameter $s = 1$, and selection function $F : x \to x - x^k$.

Remark 3.1.4. The condition (3.1.4) on μ matches the standard condition for the existence of the SLFV [BEV10]. It comes from the fact that a point $x \in \mathbb{R}^d$ is affected by a reproduction event at rate :

$$\int_{\mathbb{R}^d} \int_0^{+\infty} \mathbb{1}_{y \in \mathcal{B}(x,\mathcal{R})} \mu(d\mathcal{R}) dy = \int_0^{+\infty} V_{\mathcal{R}} \mu(d\mathcal{R}) \propto \int_0^{+\infty} \mathcal{R}^d \mu(d\mathcal{R}).$$

Remark 3.1.5. Since the density ω_M is only defined up to a Lebesgue null set, the type of individuals present in a given position $y \in \mathbb{R}^d$ cannot be uniquely defined. Therefore, even though intuitively we can first sample parental positions, and deduce parental types from ω_M, we cannot formally sample positions in order to sample parental types.

A particularly interesting feature of this model is that there exists a dual process of potential ancestors associated to it, which follows the locations of the potential ancestors of a set of individuals. In other words, the genetic diversity in a sample of the population can be determined by going *backwards in time*, and reconstructing the genealogical tree of the sample. For $k = 2$, the dual process is analogous to the Ancestral Selection Graph (ASG) [KN97; NK97], but with a spatial structure.

The k-parent ancestral process

All the new objects introduced in relation with the dual process will be defined on a new probability space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \boldsymbol{P})$. Let $\mathbf{E}[.]$ be the expectation with respect to $\boldsymbol{P}$. As before, we let μ be a σ-finite measure on $(0, +\infty)$ satisfying condition (3.1.4), and we let $\overleftarrow{\Pi}$ be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0, +\infty)$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R})$.

Let $\mathcal{M}_p(\mathbb{R}^d)$ denote the set of all finite point measures on $\mathbb{R}^d$, equipped of the topology of weak convergence. For all $\Xi = \sum_{i=1}^{l} \delta_{\xi_i} \in \mathcal{M}_p(\mathbb{R}^d)$, for all $x \in \mathbb{R}^d$ and $\mathcal{R} > 0$, we define

$$I_{x,\mathcal{R}}(\Xi) = \{i \in [\![1, l]\!] : ||x - \xi_i|| \leq \mathcal{R}\}$$
$$\text{and } S^{\mathcal{R}}(\Xi) = \{x \in \mathbb{R}^d : \exists i \in [\![1, l]\!] : ||x - \xi_i|| \leq \mathcal{R}\}.$$

In other words, $I_{x,\mathcal{R}}(\Xi)$ is the set of all the indices of the points in Ξ which are at distance at most $\mathcal{R}$ of x, while $S^{\mathcal{R}}(\Xi)$ is the set of all the points in $\mathbb{R}^d$ which are at distance at most $\mathcal{R}$ of a point of Ξ.

Definition 3.1.6. *Let $\Xi^0 \in \mathcal{M}_p(\mathbb{R}^d)$. The k-parent ancestral process $(\Xi_t)_{t \geq 0}$ associated to μ (or equivalently to $\overleftarrow{\Pi}$) and with initial condition Ξ^0 is the $\mathcal{M}_p(\mathbb{R}^d)$-valued Markov jump process defined as follows.*

- *First, we set $\Xi_0 = \Xi^0$.*
- *Then, for all $(t, x, \mathcal{R}) \in \overleftarrow{\Pi}$, if $I_{x,\mathcal{R}}(\Xi_{t-}) \neq \emptyset$ and if we write*

$$\Xi_{t-} = \sum_{i=1}^{N_{t-}} \delta_{\xi^i_{t-}},$$

we sample k points $y_1, ..., y_k$ independently and uniformly at random in $\mathcal{B}(x, \mathcal{R})$, and we set

$$\Xi_t := \sum_{i=1}^{N_{t-}} \delta_{\xi^i_{t-}} - \sum_{i \in I_{x,\mathcal{R}}(\Xi_{t-})} \delta_{\xi^i_{t-}} + \sum_{j=1}^{k} \delta_{y_j}.$$

In other words, we remove all the atoms of Ξ_{t-} sitting in $\mathcal{B}(x, \mathcal{R})$, and we add k atoms at locations that are i.i.d and uniformly distributed over the ball $\mathcal{B}(x, \mathcal{R})$.

This process is well-defined, since N_t is stochastically bounded by the number $(Y_t^k)_{t\geq 0}$ of particles in a Yule process with k children and with individual branching rate $\int_0^\infty V_{\mathcal{R}}\mu(d\mathcal{R}) < +\infty$ (see [EVY20] for a proof in the case $k = 2$, which can be generalized to the case $k \geq 2$).

The k-parent ancestral process solves a martingale problem that we now introduce. For all $F \in C_b^1(\mathbb{R})$ and $f \in \mathcal{B}(\mathbb{R}^d)$, we define the function $\Phi_{F,f} : \mathcal{M}_p(\mathbb{R}^d) \to \mathbb{R}$ by :

$$\forall \Xi \in \mathcal{M}_p(\mathbb{R}^d), \Phi_{F,f}(\Xi) = F\left(\int_{\mathbb{R}^d} f(x)\Xi(dx)\right) = F(\langle \Xi, f\rangle).$$

We now define the operator $\mathcal{G}_\mu^k$ on the set of functions of the form $\Phi_{F,f}$, which will be at the basis of the martingale problem satisfied by $(\Xi_t)_{t\geq 0}$. Let $F \in C_b^1(\mathbb{R})$ and $f \in \mathcal{B}(\mathbb{R}^d)$, then for all $\Xi = \sum_{i=1}^l \delta_{\xi_i} \in \mathcal{M}_p(\mathbb{R}^d)$, we set :

$$\mathcal{G}_\mu^k\Phi_{F,f}(\Xi) := \int_{\mathbb{R}^d}\int_0^{+\infty}\int_{\mathcal{B}(x,\mathcal{R})^k}\left[\mathbb{1}_{x\in S^{\mathcal{R}}(\Xi)} \times \frac{1}{V_{\mathcal{R}}^k} \times F\left(\langle \Xi, f\rangle - \sum_{i\in I_{x,\mathcal{R}}(\Xi)} f(x_i) + \sum_{j=1}^k f(y_j)\right)\right.$$
$$\left. -\mathbb{1}_{x\in S^{\mathcal{R}}(\Xi)} \times \frac{1}{V_{\mathcal{R}}^k} F(\langle \Xi, f\rangle)\right] dy_1...dy_k\mu(d\mathcal{R})dx$$

This operator is well defined. Indeed, for all $\Xi \in \mathcal{M}_p(\mathbb{R}^d)$, by Fubini's theorem,

$$\begin{aligned}
|\mathcal{G}_\mu^k\Phi_{F,f}(\Xi)| &\leq \int_0^{+\infty}\int_{S^{\mathcal{R}}(\Xi)}\int_{\mathcal{B}(x,\mathcal{R})^k} 2 \times \frac{1}{V_{\mathcal{R}}^k} \times ||F||_\infty dy_1...dy_k dx\mu(d\mathcal{R}) \\
&\leq 2||F||_\infty \times \int_0^{+\infty} \mathrm{Vol}(S^{\mathcal{R}(\Xi)})\mu(d\mathcal{R}) \\
&\leq 2||F||_\infty \times \Xi(\mathbb{R}^d) \times \int_0^{+\infty} V_{\mathcal{R}}\mu(d\mathcal{R}) \\
&< +\infty
\end{aligned}$$

by Condition (3.1.4).

Proposition 3.1.7. *Let $\Xi^0 \in \mathcal{M}_p(\mathbb{R}^d)$, and let $(\Xi_t)_{t\geq 0}$ be the k-parent ancestral process of initial condition Ξ^0 associated to μ. Then, for all $F \in C_b^1(\mathbb{R})$ and for all $f \in \mathcal{B}(\mathbb{R}^d)$, the process*

$$\left(\Phi_{F,f}(\Xi_t) - \Phi_{F,f}(\Xi_0) - \int_0^t \mathcal{G}_\mu^k\Phi_{F,f}(\Xi_s)ds\right)_{t\geq 0}$$

is a martingale.

The proof of Proposition 3.1.7 is a direct adaptation of the proof of Proposition 1.5 from [EVY20].

Intuitively, the k-parent ancestral process records the locations of the potential ancestors of a given sample of individuals. However, because densities are only defined up to a Lebesgue null set, it is not possible to assign uniquely a type to an individual located at $x \in \mathbb{R}^d$ looking at the value of the density at this point. Therefore, as in [EVY20], in order to give a duality relation between the k-parent SLFV and the k-parent ancestral process, we will need to consider a distribution of sampling locations, rather than fixed locations.

More specifically, for all $l \geq 1$ and $x_1, ..., x_l \in (\mathbb{R}^d)^l$, we define :

$$\Xi[x_1, ..., x_l] := \sum_{i=1}^l \delta_{x_i} \in \mathcal{M}_p(\mathbb{R}^d).$$

If Ψ is a density function on $(\mathbb{R}^d)^l$, let μ_Ψ be the law of the random point measure $\sum_{i=1}^l \delta_{X_i}$, where $(X_1, ..., X_l)$ is sampled according to Ψ. If $M \in \mathcal{M}_\lambda$ and $\Xi = \sum_{i=1}^l \delta_{x_i} \in \mathcal{M}_p(\mathbb{R}^d)$, we set :

$$D(M, \Xi) := \prod_{i=1}^{l} \omega_M(x_i).$$

Notice that for all $l \in \mathbb{N}^*$ and for all density functions Ψ on $(\mathbb{R}^d)^l$,

$$\begin{aligned}\int_{\mathcal{M}_p(\mathbb{R}^d)} D(M, \Xi)\mu_\Psi(d\Xi) &= \int_{(\mathbb{R}^d)^l} \Psi(x_1, ..., x_l) \left\{ \prod_{j=1}^{l} \omega_M(x_j) \right\} dx_1...dx_l \\ &= \int_{(\mathbb{R}^d \times \{0,1\})^l} \Psi(x_1, ..., x_l) \left\{ \prod_{j=1}^{l} \mathbb{1}_0(\kappa_j) \right\} M(dx_1, d\kappa_1)...M(dx_l, d\kappa_l)\end{aligned}$$

does not depend on the choice of a density ω_M of M.

A straightforward adaptation of the proof of Proposition 1.7 in [EVY20] to the case $k \geq 2$ leads to the following proposition.

Proposition 3.1.8. *Let $k \geq 2$. Let $M^0 \in \mathcal{M}_\lambda$, let $l \in \mathbb{N}^*$, and let Ψ be a density function on $(\mathbb{R}^d)^l$. Let μ be a σ-finite measure on $(0, +\infty)$ satisfying (3.1.4), and let $(\Xi_t)_{t\geq 0}$ be the k-parent ancestral process associated to μ. Then, any solution $(\widetilde{M}_t)_{t\geq 0}$ to the martingale problem $(\mathcal{L}^k_\mu, \delta_{M^0})$ satisfies that for all $t \geq 0$,*

$$\int_{\mathcal{M}_p(\mathbb{R}^d)} \mathbb{E}[D(M_t, \xi)|M_0 = M^0]\mu_\Psi(d\xi) = \mathbf{E}[D(M^0, \Xi_t)|\Xi_0 \sim \mu_\Psi].$$

Equivalently, for all $t \geq 0$,

$$\begin{aligned}&\mathbb{E}_{M^0}\left[\int_{(\mathbb{R}^d)^l} \Psi(x_1, ..., x_l) \left\{ \prod_{j=1}^{l} \omega_{M_t}(x_j) \right\} dx_1...dx_l\right] \\ &= \int_{(\mathbb{R}^d)^l} \Psi(x_1, ..., x_l)\mathbf{E}_{\Xi[x_1,...,x_l]}\left[\prod_{j=1}^{N_t} \omega_{M^0}(\xi^j_t)\right] dx_1...dx_l.\end{aligned}$$

3.1.2 Construction of the ∞-parent SLFV

Let us now introduce the limit of the k-parent SLFV when $k \to \infty$ somewhat informally. We will call it the ∞-parent spatial Λ-Fleming Viot process, or ∞-parent SLFV. It will be constructed rigourously in Section 3.2 using an alternative construction of the k-parent SLFV inspired by [VW15], but we now give an intuitive idea of its definition.

Let μ be a σ-finite measure on $(0, +\infty)$ satisfying Condition (3.1.4), and let Π be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0, +\infty)$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R})$. Let also $M^0 \in \mathcal{M}_\lambda$. We start the ∞-parent spatial Λ-Fleming Viot process $(M^\infty_t)_{t\geq 0}$, or ∞-parent SLFV, at $M^\infty_0 = M^0$. Then, if $(t, x, \mathcal{R}) \in \Pi$, as before, we consider that a reproduction event occurs in the ball $\mathcal{B}(x, \mathcal{R})$ at time t. However, this time we do not sample a finite number of potential parents. Instead, we look at the value of the integral

$$\int_{\mathcal{B}(x,\mathcal{R})} \left(1 - \omega_{M^\infty_{t-}}(z)\right) dz,$$

which amounts to sampling an infinite number of potential parents over the ball $\mathcal{B}(x,\mathcal{R})$ and looking at the proportion of them which are of the "existing" type (i.e, type 1).

If $\int_{\mathcal{B}(x,\mathcal{R})}\left(1-\omega_{M^\infty_{t^-}}(z)\right)dz=0$, we consider that the parent which reproduces is of type 0, and we set :

$$\forall z\in\mathcal{B}(x,\mathcal{R}),\omega_{M^\infty_t}(z)=1.$$

Note that in this case, the "parent" which reproduces was "sampled" at a location which is uniformly distributed over the ball $\mathcal{B}(x,\mathcal{R})$.

Conversely, if $\int_{\mathcal{B}(x,\mathcal{R})}\left(1-\omega_{M^\infty_{t^-}}(z)\right)dz\neq 0$, there is a non negligible number of individuals of type 1 in $\mathcal{B}(x,\mathcal{R})$. We impose that it is one of them which reproduces, and in such a way that its offspring invades the whole region. That is, we set :

$$\forall z\in\mathcal{B}(x,\mathcal{R}),\omega_{M^\infty_t}(z)=0.$$

Again, note that in our interpretation, the location of the parent which actually reproduces is uniformly distributed over the region $\{y\in\mathcal{B}(x,\mathcal{R}):\omega_{M^\infty_{t^-}}(y)=1\}$.

As for the k-parent SLFV, the ∞-parent SLFV is solution to a martingale problem. However, and in contrast with the case of the k-parent SLFV, the condition (3.1.4) on μ will not be sufficient to ensure that this solution is unique. Instead, we will need the following stronger condition.

Definition 3.1.9. *Let $a_d>0$ be such that the minimal number of d-dimensional balls of radius 1 needed to cover the boundary of an hypersphere of radius n in d dimensions is bounded from above by $a_d\times n^{d-1}$ for every $n\geq 1$. A σ-finite measure μ on $\mathbb{R}_+^*$ is said to satisfy Condition (3.1.5) if it satisfies Condition (3.1.4), and if there exists $\mathcal{R}>0$ such that*

$$\sum_{n=1}^{+\infty}\left(\int_{(n-1)\mathcal{R}}^{n\mathcal{R}}(\mathcal{R}+r)^d\mu(dr)\right)\left(a_d\times n^{d-1}+1\right)<+\infty. \tag{3.1.5}$$

Examples of σ-finite measures μ on $\mathbb{R}_+^*$ satisfying Condition (3.1.5) are the following :

1. Measures μ on $\mathbb{R}_+^*$ having a bounded support.
2. Measures μ on $\mathbb{R}_+^*$ of the form $\alpha\times(1+r)^{-3d-1}dr$, with $\alpha>0$.

We define the operator $\mathcal{L}^\infty_\mu$ on functions of the form $\Psi_{F,f}$ where $F\in C^1(\mathbb{R})$ and $f\in C_c(\mathbb{R}^d)$ in the following way. For all $M\in\mathcal{M}_\lambda$, we set :

$$\begin{aligned}\mathcal{L}^\infty_\mu\Psi_{F,f}(M):=\int_0^{+\infty}\int_{Supp^{\mathcal{R}}(f)}\Bigg[&\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_M(z))\,dz\right)\times F(\langle\Theta^+_{x,\mathcal{R}}(\omega_M),f\rangle)\\&+\left(1-\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_M(z))\,dz\right)\right)\times F(\langle\Theta^-_{x,\mathcal{R}}(\omega_M),f\rangle)\\&-F(\langle\omega_M,f\rangle)\Bigg]dx\mu(d\mathcal{R}).\end{aligned}$$

Note that if $\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_M(z))\,dz\right)=1$, then for all $y\in\mathcal{B}(x,\mathcal{R})$ except possibly on a Lebesgue null set,

$$\Theta^+_{x,\mathcal{R}}(\omega_M)(y)=\omega_M(y).$$

In other words, the ball $\mathcal{B}(x,\mathcal{R})$ is already completely void of "existing" individuals, and filling it with "ghost" individuals does not change anything. Therefore, we also have

$$\mathcal{L}_\mu^\infty \Psi_{F,f}(M) = \int_0^{+\infty} \int_{Supp^{\mathcal{R}}(f)} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_M(z))\, dz\right)\right) \times \Big[F(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f\rangle) - F(\langle \omega_M, f\rangle)\Big] dx\mu(d\mathcal{R}).$$

In Section 3.5, we show that this operator is well-defined, even if μ satisfies Condition (3.1.4) rather than Condition (3.1.5). If μ satisfies Condition (3.1.5), then the associated martingale problem can be used to define and fully characterize the ∞-parent SLFV. If μ satisfies only Condition (3.1.4), then the ∞-parent SLFV is still solution to the martingale problem, but we no longer know whether this solution is unique, as stated in the following theorem. Therefore, we will provide in Section 3.2 a construction of the ∞-parent SLFV which does not rely on the martingale problem, and works even if μ only satisfies Condition (3.1.4).

Theorem 3.1.10. *Let $\omega : \mathbb{R}^d \to \{0,1\}$, let $M^0 \in \mathcal{M}_\lambda$, and let μ be a σ-finite measure on $(0,+\infty)$ satisfying Condition (3.1.4). Then, the ∞-parent SLFV with initial condition M^0 associated to μ defined in Section 3.2.2 is a solution to the martingale problem for $(\mathcal{L}_\mu^\infty, \delta_{M^0})$.*

Moreover, if μ satisfies Condition (3.1.5), the martingale problem associated to $(\mathcal{L}_\mu^\infty, \delta_{M^0})$ is well-posed, and the ∞-parent SLFV with initial condition M^0 associated to μ is the unique solution to it in $D_{\mathcal{M}_\lambda}[0,+\infty)$.

The proof of Theorem 3.1.10 can be found at the end of Section 3.4.3.

3.1.3 Dual of the ∞-parent SLFV

As for the k-parent SLFV, the ∞-parent SLFV also has a dual process of potential ancestors.

Let $\mathcal{E}^c$ be the set of Lebesgue measurable, closed and connected subsets of $\mathbb{R}^d$ whose Lebesgue measure is finite and strictly positive. Let $\mathcal{E}^{cf}$ be the set of all finite unions of elements of $\mathcal{E}^c$. If $E \in \mathcal{E}^{cf}$ can be written as $E = \cup_{i=1}^l E^i$ where for all $1 \le i \le l$, $E^i \in \mathcal{E}^c$, we let $m(E) = m(E^1,...,E^l)$ be the measure on $\mathbb{R}^d$ defined by $m(E)(dx) := \mathbb{1}_{x\in E} dx$, and we set :

$$\mathcal{M}^{cf} := \{m(E) : E \in \mathcal{E}^{cf}\}.$$

Definition 3.1.11 (∞-parent ancestral process)**.** *Let μ be a σ-finite measure on $(0,+\infty)$ satisfying Condition (3.1.5). Let $\overleftarrow{\Pi}$ be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^d \times (0,+\infty)$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R})$, defined on the probability space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \boldsymbol{P})$.*
Let $\Xi^0 = m(E_0^1,...,E_0^l) \in \mathcal{M}^{cf}$. Then, the $\mathcal{M}^{cf}$-valued ∞-parent ancestral process $(\Xi_t^\infty)_{t\ge 0}$ with initial condition Ξ^0 associated to μ (or equivalently to $\overleftarrow{\Pi}$) is defined in the following way.

First, we set $\Xi_0^\infty = \Xi^0$. Then, if for all $t \ge 0$, we write Ξ_t^∞ as

$$\Xi_t^\infty = m(E_t),$$

then for all $(t,x,\mathcal{R}) \in \overleftarrow{\Pi}$, if $E_{t-} \cap \mathcal{B}(x,\mathcal{R})$ has a non zero Lebesgue measure,

$$\Xi_t^\infty = m(E_{t-} \cup \mathcal{B}(x,\mathcal{R})).$$

Moreover, this process is Markovian.

Since the initial condition $\Xi^0 \in \mathcal{M}^{cf}$ and since a (closed) ball is added to its support whenever the process jumps, it is clearly $\mathcal{M}^{cf}$-valued. We will show that this process is well-defined and Markovian in Section 3.3.

Remark 3.1.12. Note that the case $E_{t-} \cap \mathcal{B}(x, \mathcal{R}) = \mathcal{B}(x, \mathcal{R})$ is equivalent to $E_{t-} \cup \mathcal{B}(x, \mathcal{R}) = E_{t-}$, and hence does not correspond to a visible jump of $(\Xi_t^\infty)_{t\geq 0}$.

For all $M \in \mathcal{M}_\lambda$ with density ω and for all $\Xi = m(E) \in \mathcal{M}^{cf}$, we set :

$$\widetilde{D}(M, \Xi) := \delta_0 \left(\int_E (1 - \omega(x))\, dx \right).$$

If we know the value of $\widetilde{D}(M, \Xi)$ for all $\Xi \in \mathcal{M}^{cf}$, since ω is $\{0, 1\}$-valued, we know the value of ω everywhere up to a Lebesgue null set, and so we have completely characterized M. Therefore, the following duality result shows that the solution to the martingale problem associated to $\mathcal{L}_\mu^\infty$ is unique.

Proposition 3.1.13. *Let μ be a σ-finite measure on $(0, +\infty)$ satisfying Condition (3.1.5). Let $M^0 \in \mathcal{M}_\lambda$, and let $(M_t^\infty)_{t\geq 0}$ be a solution to the martingale problem associated to $(\mathcal{L}_\mu^\infty, \delta_{M^0})$. Then, for all $t \geq 0$ and for all $E^0 \in \mathcal{M}^{cf}$,*

$$\mathbb{E}_{M^0} \left[\widetilde{D}(M_t^\infty, m(E^0)) \right] = \mathbf{E}_{m(E^0)} \left[\widetilde{D}(M^0, \Xi_t^\infty) \right],$$

where (Ξ_t^∞) is the ∞-parent ancestral process of initial condition $m(E^0)$ associated to μ. Equivalently, for every $t \geq 0$, if ω_t and ω_0 are $\{0, 1\}$-valued densities of M_t^∞ and M^0,

$$\mathbb{E} \left[\delta_0 \left(\int_{E^0} (1 - \omega_t(x))\, dx \right) \right] = \mathbf{E}_{m(E^0)} \left[\delta_0 \left(\int_{E_t} (1 - \omega_0(x))\, dx \right) \right].$$

The proof of Proposition 3.1.13 can be found in Section 3.4.3.

3.1.4 Structure of the paper

In Section 2, we construct the ∞-parent SLFV rigorously, by introducing a coupling between a sequence of k-parent SLFV processes with the same initial conditions. We also show the first part of Theorem 3.1.10, i.e, that the ∞-parent SLFV is a solution to the martingale problem associated to $\mathcal{L}_\mu^\infty$. In Section 3, we first demonstrate that the ∞-parent ancestral process is well defined, and we then show that it can be characterized as the unique solution to a specific martingale problem. Section 3.4 is devoted to the proof of the duality relation between the ∞-parent SLFV and the ∞-parent ancestral process stated in Proposition 3.1.13. The second part of Theorem 3.1.10 is then a direct consequence of Proposition 3.1.13. Section 3.5 contains technical lemmas used throughout the paper.

3.2 The ∞-parent SLFV

3.2.1 Alternative construction of the k-parent SLFV

In order to construct the ∞-parent SLFV rigorously, we start by introducing an alternative construction of the k-parent SLFV, based on a variant of its dual. It relies on the sampling of parental locations *along with* reproduction events, and is an adaptation of the concept of *parental skeleton* presented in Section 2.3.1 of [VW15].

In all that follows, let μ be a σ-finite measure on $(0, +\infty)$ satisfying Condition (3.1.4). Let $U = \mathcal{B}(0, 1)^{\mathbb{N}}$, and let $\tilde{u}$ be the law of a sequence of i.i.d random variables $(\mathcal{P}_n)_{n\geq 1}$ uniformly distributed over $\mathcal{B}(0, 1)$. We will call an element of U a *sequence of potential parents*. Let us now extend the Poisson point process Π considered earlier by adding to each event a countable sequence of

locations of potential parents. Indeed, let Π^c be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0, +\infty) \times U$ with intensity

$$dt \otimes dx \otimes \mu(d\mathcal{R}) \otimes \tilde{u}(d(p_n)_{n\geq 1}).$$

Then for all $(t, x, \mathcal{R}, (p_n)_{n\geq 1}) \in \Pi^c$,

- as before, t can be interpreted as the time at which the reproduction event occurs, and we can see $\mathcal{B}(x, \mathcal{R})$ as being the area affected by the reproduction event.
- For all $n \geq 1$, $x + \mathcal{R} \times p_n$ is uniformly distributed over the ball $\mathcal{B}(x, \mathcal{R})$, and can be interpreted as the location of the n-th potential parent sampled, if at least n potential parents have to be sampled.

We start by defining the variant of the k-parent ancestral process, on which the alternative construction of the k-parent SLFV is based.

Definition 3.2.1 (Quenched k-parent ancestral process)**.** *Let $k \geq 2$, let $\Xi^0 \in M_p(\mathbb{R}^d)$, and let $\tilde{t} \geq 0$. The k-parent ancestral process $(\Xi_{k,t}^{\Pi^c,\tilde{t},\Xi^0})_{t\geq 0}$ associated to Π^c, started at time $\tilde{t}$ and with initial condition Ξ^0 is the $\mathcal{M}_p(\mathbb{R}^d)$-valued Markov jump process defined as follows.*

- *First, we set $\Xi_{k,0}^{\Pi^c,\tilde{t},\Xi^0} = \Xi^0$.*
- *Then, for all $(t, x, \mathcal{R}, (p_n)_{n\geq 1}) \in \Pi^c$ such that $t \leq \tilde{t}$, recalling that for $\Xi = \sum_{i=1}^{l} \delta_{\xi_i} \in \mathcal{M}_p(\mathbb{R}^d)$, $I_{x,\mathcal{R}}(\Xi) = \{i \in [\![1, l]\!] : ||x - \xi_i|| \leq \mathcal{R}\}$, if*

  $$I_{x,\mathcal{R}}(\Xi_{k,(\tilde{t}-t)_-}^{\Pi^c,\tilde{t},\Xi^0}) \neq \emptyset,$$

  then for all $1 \leq l \leq k$, we set

  $$y_l := x + \mathcal{R} \times p_l$$

  $$\text{and} \qquad \Xi_{k,\tilde{t}-t}^{\Pi^c,\tilde{t},\Xi^0} := \Xi_{k,(\tilde{t}-t)-}^{\Pi^c,\tilde{t},\Xi^0} - \sum_{x' \in I_{x,\mathcal{R}}(\Xi_{k,(\tilde{t}-t)_-}^{\Pi^c,\tilde{t},\Xi^0})} \delta_{x'} + \sum_{l=1}^{k} \delta_{y_l}.$$

It is straightforward to check that this process has the same distribution as the k-parent ancestral process associated to μ and with initial condition Ξ^0. Its interest is twofold. First, conditionally on Π^c, $(\Xi_{k,t}^{\Pi^c,\tilde{t},\Xi^0})_{t\geq 0}$ is completely deterministic. Moreover, if for all $\Xi = \sum_{i=1}^{l} \delta_{x_i} \in \mathcal{M}_p(\mathbb{R}^d)$, we denote the set of atoms of Ξ by

$$A(\Xi) := \{x_i : i \in [\![1, l]\!]\},$$

then the process satisfies the following property, which will be useful in the coupling that we will introduce later.

Lemma 3.2.2. *Let $2 \leq k \leq k'$, let $\Xi^0 \in \mathcal{M}_p(\mathbb{R}^d)$, let $\tilde{t} \geq 0$, and let Π^c be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0, +\infty) \times U$ with intensity $dt \otimes dx \otimes \mu(d\mathcal{R}) \otimes \tilde{u}(d(p_n)_{n\geq 1})$.*

Then, for all $t \geq 0$,

$$A(\Xi_{k,t}^{\Pi^c,\tilde{t},\Xi^0}) \subseteq A(\Xi_{k',t}^{\Pi^c,\tilde{t},\Xi^0}).$$

In particular, for all $t \geq 0$ and $x \in \mathbb{R}^d$,

$$A(\Xi_{k,t}^{\Pi^c,\tilde{t},\delta_x}) \subseteq A(\Xi_{k',t}^{\Pi^c,\tilde{t},\delta_x}).$$

Remark 3.2.3. Since $A(\Xi)$ is a set, if there exists $i \neq j$ such that $x_i = x_j$, then x_i appears only once in $A(\Xi)$.

Intuitively, the idea behind this lemma is the following. Since the coupled k-parent and k'-parent ancestral processes are based on the same extended Poisson point process of reproduction events, their evolutions are determined by the same reproduction events. Moreover, since $k' \geq k$, all the potential parents which are involved in the dynamics of the k-parent ancestral process are also potential parents for the k'-ancestral process. Therefore, we can consider that the k-parent ancestral process is embedded in the k'-parent ancestral process.

We now introduce an alternative way of constructing the k-parent SLFV, by associating it to the extended Poisson point process Π^c.

Definition 3.2.4 (Quenched k-parents SLFV)**.** *Let $k \geq 2$, and let $\omega : \mathbb{R}^d \to \{0,1\}$ be a measurable function. The k-parent SLFV $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0}$ associated to Π^c and of initial density ω is the $\mathcal{M}_\lambda$-valued Markov process defined as follows.*

- *First, we set $\omega_{k,0}^{\Pi^c,\omega} = \omega$.*
- *Then, for all $t \geq 0$ and for all $x \in \mathbb{R}^d$, we set*

$$\omega_{k,t}^{\Pi^c,\omega}(x) := \prod_{y \in A(\Xi_{k,t}^{\Pi^c,t,\delta_x})} \omega(y). \tag{3.2.1}$$

- *We conclude by setting for all $t \geq 0$,*

$$M_{k,t}^{\Pi^c,\omega} := ((\omega_{k,t}^{\Pi^c,\omega}(x)\delta_0(d\kappa) + (1 - \omega_{k,t}^{\Pi^c,\omega}(x))\delta_1(d\kappa))dx.$$

$(\omega_{k,t}^{\Pi^c,\omega})_{t\geq 0}$ will be called the density of the k-parent SLFV associated to Π^c and of initial condition ω.

Note that $\omega_{k,t}^{\Pi^c,\omega}(x)$ in Eq. 3.2.1 is thus equal to 1 if and only if all potential ancestors at time 0 of the individuals at x at time t are of type 0, i.e are all ghosts.

We show below that this process corresponds to another way of constructing the k-parent SLFV using the parental skeleton, and in particular, that $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0} \in D_{M_\lambda}[0,+\infty)$. This alternative construction will allow us to couple SLFV processes with different numbers of potential parents, using the same Poisson process. However, even though it is possible to define the k-parent SLFV for an *initial condition* $M \in \mathcal{M}_\lambda$ instead of an initial density ω of M, this coupling can only be used if all processes are constructed using the same initial density.

Lemma 3.2.5. *Under the notation of Definition 3.2.4, $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0} \in D_{M_\lambda}[0,+\infty)$.*

Proof. In order for the process to have a chance to correspond to the k-parent SLFV, we first need to check that

$$(M_{k,t}^{\Pi^c,\omega})_{t\geq 0} \in D_{M_\lambda}[0,+\infty).$$

Let $t \geq 0$. Since ω is $\{0,1\}$-valued, by definition $\omega_{k,t}^{\Pi^c,\omega}$ is $\{0,1\}$-valued. Moreover, the values taken by ω are changed over balls of the form $\mathcal{B}(x,\mathcal{R})$ in order to compute $\omega_{k,t}^{\Pi^c,\omega}$, and since any compact area is affected by reproduction events at a finite rate, the value of $\omega_{k,t}^{\Pi^c,\omega}$ over any given compact set only depends on the initial condition ω and on a (a.s.) finite number of reproduction events. Therefore, as ω is measurable, $\omega_{k,t}^{\Pi^c,\omega}$ is measurable as well, and we obtain that for all $t \geq 0$, $M_{k,t}^{\Pi^c,\omega} \in \mathcal{M}_\lambda$.

We now show that the process is a.s. càdlàg (for the topology of vague convergence). That is, we want to show that

1. $\lim_{s\uparrow t} M_{k,s}^{\Pi^c,\omega}$ exists,

2. $\lim_{s\downarrow t} M_{k,s}^{\Pi^c,\omega}$ exists and is equal to $M_{k,t}^{\Pi^c,\omega}$.

In order to do so, let $(f_n)_{n\in\mathbb{N}} \in C_c(\mathbb{R}^d)$ be a convergence determining class. Then, it is sufficient to show that

1. For all $n \in \mathbb{N}$, $\lim_{s\uparrow t}\langle \omega_{k,s}^{\Pi^c,\omega}, f_n\rangle$ exists a.s.,

2. For all $n \in \mathbb{N}$, $\lim_{s\downarrow t}\langle \omega_{k,s}^{\Pi^c,\omega}, f_n\rangle$ exists and is equal to $\langle \omega_{k,t}^{\Pi^c,\omega}, f_n\rangle$ a.s.

Therefore, let $n \in \mathbb{N}$. We set

$$t_{f-}^{(n)} := \sup\left\{t' < t : \exists \mathcal{R} > 0, \exists x \in \mathrm{Supp}^{\mathcal{R}}(f), \exists (p_n)_{n\geq 1} \in U, (t', x, \mathcal{R}, (p_n)_{n\geq 1}) \in \Pi^c\right\}$$
$$\text{and} \quad t_{f+}^{(n)} := \inf\left\{t' > t : \exists \mathcal{R} > 0, \exists x \in \mathrm{Supp}^{\mathcal{R}}(f), \exists (p_n)_{n\geq 1} \in U, (t', x, \mathcal{R}, (p_n)_{n\geq 1}) \in \Pi^c\right\}.$$

As there exists $C^{(n)} > 0$ such that for all $\mathcal{R} > 0$,

$$\mathrm{Vol}(\mathrm{Supp}^{\mathcal{R}}(f)) \leq C^{(n)}(\mathcal{R}^d \vee 1),$$

the support of f is affected by reproduction events at rate

$$\begin{aligned}\int_0^\infty \mathrm{Vol}(\mathrm{Supp}^{\mathcal{R}}(f)) &\leq C^{(n)} \int_0^\infty (\mathcal{R}^d \vee 1)(d\mathcal{R}) \\ &< +\infty\end{aligned}$$

since satisfies Condition (3.1.4). Therefore, $t_{f-}^{(n)} < t$ and $t_{f+}^{(n)} > t$ a.s., and we obtain

$$\lim_{s\uparrow t}\langle \omega_{k,s}^{\Pi^c,\omega}, f_n\rangle = \langle \omega_{k,t_{f-}^{(n)}}^{\Pi^c,\omega}, f_n\rangle \quad \text{a.s.}$$
$$\text{and} \quad \langle \omega_{k,s}^{\Pi^c,\omega}, f_n\rangle = \langle \omega_{k,t}^{\Pi^c,\omega}, f_n\rangle \quad \text{a.s.},$$

allowing us to conclude. □

Lemma 3.2.6. *Under the notation of Definition 3.2.4, $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0}$ has the same distribution as the k-parent SLFV associated to μ and with initial density ω.*

Proof. In order to show this result, we follow the following two steps:

1. We show that $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0}$ is Markovian.

2. We show that it satisfies the duality relation stated in Proposition 3.1.8.

Indeed, a direct adaptation of the proof of Theorem 1.2 from [EVY20] implies that the k-parent SLFV (which also satisfies this duality relation by Proposition 3.1.8) is then equal in distribution to $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0}$.

First, we show that $(M_{k,t}^{\Pi^c,\omega})_{t\geq 0}$ is Markovian. In order to do so, we observe that it is sufficient to show that $(\omega_{k,t}^{\Pi^c,\omega})_{t\geq 0}$ is Markovian (as a process taking its values in the set of measurable functions from $\mathbb{R}^d$ to $\{0,1\}$, and for the filtration given by the extended Poisson point process Π^c). The result is then a direct consequence of Lemma 3.5.10 from Section 3.5.

Then, we set $M^0 = M^{\Pi^c,\omega}_{k,0}$. Let $l \in \mathbb{N}^*$, let Ψ be a density function on $(\mathbb{R}^d)^l$, and let $t \geq 0$. Then,

$$\begin{aligned}
&\mathbb{E}_{M^0}\left[\int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\left(\prod_{j=1}^{l} \omega^{\Pi^c,\omega}_{k,t}(x_j)\right) dx_1...dx_l\right] \\
&= \int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\mathbb{E}_{M^0}\left[\prod_{j=1}^{l} \omega^{\Pi^c,\omega}_{k,t}(x_j)\right] dx_1...dx_l \\
&= \int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\mathbb{E}_{M^0}\left[\prod_{j=1}^{l} \prod_{y\in A(\Xi^{\Pi^c,t,\delta_{x_j}}_{k,t})} \omega(y)\right] dx_1...dx_l \\
&= \int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\mathbb{E}_{M^0}\left[\prod_{y\in A(\Xi^{\Pi^c,t,\sum_{j=1}^{l}\delta_{x_j}}_{k,t})} \omega(y)\right] dx_1...dx_l \\
&= \int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\mathbf{E}_{\Xi[x_1,...,x_l]}\left[\prod_{y\in A(\Xi_t)} \omega(y)\right] dx_1...dx_l,
\end{aligned}$$

with $(\Xi_t)_{t\geq 0}$ the k-parent ancestral process associated to μ with initial condition $\Xi[x_1,...,x_l]$. We used the definition of the quenched k-parent SLFV to pass from line 2 to line 3, the fact that ω is $\{0,1\}$-valued to pass from line 3 to line 4, and the observation that the k-parent ancestral process and its quenched version have the same distribution to conclude.

Writing $\Xi_t = \sum_{j=1}^{N_t} \xi^j_t$, we obtain

$$\begin{aligned}
&\mathbb{E}_{M^0}\left[\int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\left(\prod_{j=1}^{l} \omega^{\Pi^c,\omega}_{k,t}(x_j)\right) dx_1...dx_l\right] \\
&= \int_{(\mathbb{R}^d)^l} \Psi(x_1,...,x_l)\mathbf{E}_{\Xi[x_1,...,x_l]}\left[\prod_{j=1}^{N_t} \omega(\xi^j_t)\right] dx_1...dx_l.
\end{aligned}$$

This concludes the proof. □

This lemma has two direct consequences. First, $(M^{\Pi^c,\omega}_{k,t})_{t\geq 0}$ is Markovian. Moreover, since this process is $\mathcal{M}_\lambda$-valued, we have proved the second part of Theorem 3.1.1, that is, that the unique solution to the martingale problem characterizing the k-parent SLFV is $\mathcal{M}_\lambda$-valued.

The interest of the coupling lies in the fact that given a sequence of coupled k-parent SLFV constructed using the same extended Poisson point process Π^c, their corresponding densities, as constructed in Definition 3.2.4, satisfy the following property.

Lemma 3.2.7. *Let $2 \leq k < k'$, and let $\omega : \mathbb{R}^d \to \{0,1\}$ be a measurable function. Let Π^c be a Poisson point process on $\mathbb{R}\times\mathbb{R}^d\times(0,+\infty)\times U$ with intensity $dt\otimes dx\otimes\mu(d\mathcal{R})\otimes\tilde{u}((p_n)_{n\geq 1})$.*

Then, for all $t \geq 0$ and $x \in \mathbb{R}^d$,

$$\omega^{\Pi^c,\omega}_{k',t}(x) \leq \omega^{\Pi^c,\omega}_{k,t}(x).$$

In particular, for all $t \geq 0$ and $x \in \mathbb{R}^d$, $(\omega^{\Pi^c,\omega}_{k,t}(x))_{k\geq 2}$ converges to some $\omega^\infty_t(x) \in \{0,1\}$ as $k \to +\infty$.

Proof. Let $t \geq 0$ and $x \in \mathbb{R}^d$. By Lemma 3.2.2,

$$A(\Xi_{k,t}^{\Pi^c,t,\delta_x}) \subseteq A(\Xi_{k',t}^{\Pi^c,t,\delta_x}).$$

Therefore, as ω is $\{0,1\}$-valued,

$$\begin{aligned}\omega_{k',t}^{\Pi^c,\omega}(x) &= \prod_{y \in A(\Xi_{k',t}^{\Pi^c,t,\delta_x})} \omega(y) \\ &\leq \prod_{y \in A(\Xi_{k,t}^{\Pi^c,t,\delta_x})} \omega(y) \\ &\leq \omega_{k,t}^{\Pi^c,\omega}(x).\end{aligned}$$

The second part of the lemma is a consequence of the fact that $(\omega_{k,t}^{\Pi^c,\omega}(x))_{k\geq 2}$ is a non-increasing $\{0,1\}$-valued sequence. □

3.2.2 Definition of the ∞-parent SLFV

We can now define the ∞-parent SLFV.

Definition 3.2.8. *Let $M^0 \in \mathcal{M}_\lambda$ with density $\omega : \mathbb{R}^d \to \{0,1\}$. The ∞-parent spatial Λ-Fleming Viot process, or ∞-parent SLFV, with initial density ω associated to the extended Poisson point process Π^c is the $\mathcal{M}_\lambda$-valued process $(M_t^\infty)_{t\geq 0}$ defined the following way.*

First, we set $M_0^\infty = M^0$. Then, for all $t \geq 0$ and $x \in \mathbb{R}^d$, we set

$$\omega_t^\infty(x) := \lim_{k \to +\infty} \omega_{k,t}^{\Pi^c,\omega}(x)$$

and we set

$$M_t^\infty(dx, d\kappa) := (\omega_t^\infty(x)\delta_0(d\kappa) + (1 - \omega_t^\infty(x))\delta_1(d\kappa))dx.$$

Π^c will be called the associated extended Poisson point process, and $(\omega_t^\infty)_{t\geq 0}$ will be called the density of the ∞-parent SLFV associated to Π^c and of initial density ω.

In its more general form, the ∞-parent SLFV is defined for an initial condition $M^0 \in \mathcal{M}_\lambda$ and a σ-finite measure μ. However, we construct it using a density ω of M^0, and an extended Poisson point process Π^c, and in the following, we will need both the initial density and the extended Poisson process used in order to prove some properties satisfied by the ∞-parent SLFV. Therefore, we considered two complementary definitions of the process, one based on the initial condition and the measure μ, and the other one based on the initial density and the extended Poisson point process, both definitions corresponding to the same process (in law). In the following, we will use one or the other of the two definitions, depending on whether the initial density and extended Poisson point process used to construct the process are needed or not.

As in the proof of Definition 3.2.4, we can show that $(M_t^\infty)_{t\geq 0} \in D_{\mathcal{M}_\lambda}[0, +\infty)$.

Lemma 3.2.9. *Under the notation of Definition 3.2.8, $(M_t^\infty)_{t\geq 0}$ is Markovian.*

Proof. First, notice that the definition of $(M_t^\infty)_{t\geq 0}$ implies that we only need to demonstrate that $(\omega_t^\infty)_{t\geq 0}$ is Markovian (where $(\omega_t^\infty)_{t\geq 0}$ is considered as a process taking its values in the space of all measurable functions from $\mathbb{R}^d$ to $\{0,1\}$, and considering the filtration given by the associated extended Poisson point process Π^c).

Let $0 \leq s \leq t$ and let $x \in \mathbb{R}^d$. Our goal is to show that

$$\omega_t^\infty(x) = \lim_{\tilde{k}\to+\infty} \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{\tilde{k},t-s}\right)} \omega_s^\infty(x').$$

Indeed, if this result is true, since $A\left(\Xi^{\Pi^c,t,\delta_x}_{\tilde{k},t-s}\right)$ depends on events occuring during the interval $[s,t]$, it is independent from $(\omega_{s'}^\infty)_{0\leq s'\leq s}$ and we can conclude.

By definition of the ∞-parent SLFV,

$$\omega_t^\infty(x) = \lim_{k\to+\infty} \omega_{k,t}^{\Pi^c,\omega}(x).$$

Using Lemma 3.5.10 from Section 3.5, we obtain

$$\omega_t^\infty(x) = \lim_{k\to+\infty} \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{k,t-s}\right)} \omega_{k,s}^{\Pi^c,\omega}(x').$$

Let $\tilde{k} \geq 2$. By Lemma 3.2.7 and since for all $k \geq 2$, $\omega_{k,s}^{\Pi^c,\omega}$ is $\{0,1\}$-valued,

$$\begin{aligned}
\omega_t^\infty(x) &\leq \lim_{k\to+\infty} \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{\tilde{k},t-s}\right)} \omega_{k,s}^{\Pi^c,\omega}(x') \\
&\leq \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{\tilde{k},t-s}\right)} \lim_{k\to+\infty} \omega_{k,s}^{\Pi^c,\omega}(x') \\
&\leq \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{\tilde{k},t-s}\right)} \omega_s^\infty(x').
\end{aligned}$$

Here we used Lemma 3.5.11 to pass from the first to the second line, and the definition of the ∞-parent SLFV to pass from the second to the third line.

Since this is true for all $\tilde{k} \geq 2$,

$$\omega_t^\infty(x) \leq \lim_{\tilde{k}\to+\infty} \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{\tilde{k},t-s}\right)} \omega_s^\infty(x').$$

Then, starting back from the equation

$$\omega_t^\infty(x) = \lim_{k\to+\infty} \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{k,t-s}\right)} \omega_{k,s}^{\Pi^c,\omega}(x'),$$

as for all $x \in \mathbb{R}^d$, $\left(\omega_{k,s}^{\Pi^c,\omega}(x')\right)_{k\geq 2}$ is decreasing, we obtain that

$$\omega_t^\infty(x) \geq \lim_{k\to+\infty} \prod_{x' \in A\left(\Xi^{\Pi^c,t,\delta_x}_{k,t-s}\right)} \omega_s^\infty(x')$$

and we can conclude. □

3.2.3 Characterization via a martingale problem

Let $M^0 \in \mathcal{M}_\lambda$ with density $\omega : \mathbb{R}^d \to \{0,1\}$. We recall that the operator $\mathcal{L}_\mu^\infty$ is defined by

$$\mathcal{L}_\mu^\infty \Psi_{F,f}(M) = \int_0^{+\infty} \int_{Supp^{\mathcal{R}}(f)} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_M(z))\,dz\right)\right) \times \Big[F(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f\rangle) - F(\langle \omega_M, f\rangle)\Big] dx\mu(d\mathcal{R}).$$

The goal of this section is to demonstrate the following result, which is also the first part of Theorem 3.1.10.

Proposition 3.2.10. *Let $(M_t^\infty)_{t\geq 0}$ be the ∞-parent SLFV with initial density ω, associated to Π^c. Then, for all $F \in C^1(\mathbb{R})$ and $f \in C_c(\mathbb{R}^d)$,*

$$\left(\Psi_{F,f}(M_t) - \Psi_{F,f}(M_0) - \int_0^t \mathcal{L}_\mu^\infty \Psi_{F,f}(M_s) ds\right)_{t\geq 0}$$

is a martingale.

In other words, $(M_t^\infty)_{t\geq 0}$ is a solution of the martingale problem $(\mathcal{L}_\mu^\infty, \delta_{M_0^\infty})$, but this solution is not necessarily unique. In fact, we will show in Section 3.4 that this solution is unique when μ satisfies the stronger Condition (3.1.5), but the question of uniqueness when μ does not satisfy Condition (3.1.5) remains open.

We start by justifying why the operator $\mathcal{L}_\mu^\infty$ is a suitable candidate for an operator characterizing the limit $k \to +\infty$ of the k-parent SLFV.

Lemma 3.2.11. *Let $\omega : \mathbb{R}^d \to [0,1]$, and let $x \in \mathbb{R}$. Then, for all $\mathcal{R} > 0$,*

$$\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega(z))\,dz\right) = \lim_{k\to+\infty} \frac{1}{V_{\mathcal{R}}^k} \int_{\mathcal{B}(x,\mathcal{R})^k} \left(\prod_{j=1}^k \omega(y_j)\right) dy_1 ... dy_k.$$

Proof. For all $k \geq 2$,

$$\frac{1}{V_{\mathcal{R}}^k} \int_{\mathcal{B}(x,\mathcal{R})^k} \left[\prod_{j=1}^k \omega(y_j)\right] dy_1 ... dy_k = \left(\frac{1}{V_{\mathcal{R}}} \int_{\mathcal{B}(x,\mathcal{R})} \omega(y) dy\right)^k.$$

As $V_{\mathcal{R}}^{-1} \int_{\mathcal{B}(x,\mathcal{R})} \omega(y)dy \in [0,1]$,

$$\begin{aligned} \lim_{k\to+\infty} \frac{1}{V_{\mathcal{R}}^k} \int_{\mathcal{B}(x,\mathcal{R})^k} \left[\prod_{j=1}^k \omega(y_j)\right] dy_1 ... dy_k = 1 \\ \iff \frac{1}{V_{\mathcal{R}}} \int_{\mathcal{B}(x,\mathcal{R})} \omega(y) dy = 1 \\ \iff \int_{\mathcal{B}(x,\mathcal{R})} (1-\omega(z))\,dz = 0. \end{aligned}$$

Moreover,

$$\lim_{k\to+\infty} \frac{1}{V_{\mathcal{R}}^k} \int_{\mathcal{B}(x,\mathcal{R})^k} \left[\prod_{j=1}^{k} \omega(y_j)\right] dy_1...dy_k = 0$$
$$\iff \frac{1}{V_{\mathcal{R}}} \int_{\mathcal{B}(x,\mathcal{R})} \omega(y)dy < 1$$
$$\iff \int_{\mathcal{B}(x,\mathcal{R})} (1-\omega(z))\, dz > 0,$$

and we can conclude. □

Let $F \in C^1(\mathbb{R})$ and $f \in C_c(\mathbb{R}^d)$. For all $M \in \mathcal{M}_\lambda$,

$$|F(\langle \omega_M, f\rangle)| \leq \max\{|F(x)| : x \in [-\mathrm{Vol}(Supp(f)), \mathrm{Vol}(Supp(f))]\}, \tag{3.2.2}$$

which means in particular that for all $x \in \mathbb{R}^d$ and for all $\mathcal{R} > 0$,

$$|F(\langle \Theta^+_{x,\mathcal{R}}(\omega_M), f\rangle)| \leq \max\{|F(x)| : x \in [-\mathrm{Vol}(Supp(f)), \mathrm{Vol}(Supp(f))]\}$$
and $$|F(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f\rangle)| \leq \max\{|F(x)| : x \in [-\mathrm{Vol}(Supp(f)), \mathrm{Vol}(Supp(f))]\}.$$

Therefore, a direct consequence of the dominated convergence theorem is the following lemma.

Lemma 3.2.12. *Let $M \in \mathcal{M}_\lambda$, and let $(M_n)_{n\in\mathbb{N}} \in \mathcal{M}_\lambda$ be such that M_n converges vaguely to M. Then, for all $x \in \mathbb{R}^d$ and for all $\mathcal{R} > 0$,*

$$F(\langle \omega_{M_n}, f\rangle) \xrightarrow[n\to+\infty]{} F(\langle \omega_M, f\rangle)$$
$$F(\langle \Theta^+_{x,\mathcal{R}}(\omega_{M_n}), f\rangle) \xrightarrow[n\to+\infty]{} F(\langle \Theta^+_{x,\mathcal{R}}(\omega_M), f\rangle)$$
$$F(\langle \Theta^-_{x,\mathcal{R}}(\omega_{M_n}), f\rangle) \xrightarrow[n\to+\infty]{} F(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f\rangle).$$

In contrast with $\mathcal{L}^k_\mu \Psi_{F,f}$, the function $\mathcal{L}^\infty_\mu \Psi_{F,f}$ is not continuous on $\mathcal{M}_\lambda$. However, we have the following result.

Lemma 3.2.13. *Let $M \in \mathcal{M}_\lambda$, and $(M_n)_{n\in\mathbb{N}} \in \mathcal{M}_\lambda$ such that M_n converges to M in the topology of vague convergence. Assume that there exists a density ω of M and densities ω_n of M_n for all $n \in \mathbb{N}$ such that :*

$$\forall n \in \mathbb{N}, \forall z \in \mathbb{R}^d, \omega(z) \leq \omega_n(z).$$

Then,

$$\lim_{n\to+\infty} \mathcal{L}^\infty_\mu \Psi_{F,f}(M_n) = \mathcal{L}^\infty_\mu \Psi_{F,f}(M).$$

Proof. First, since $(M_n)_{n\in\mathbb{N}}$ converges vaguely to M, by Lemma 3.2.12, for all $\mathcal{R} > 0$ and for all $x \in Supp^{\mathcal{R}}(f)$,

$$F(\langle \Theta^-_{x,\mathcal{R}}(\omega_n), f\rangle) \xrightarrow[n\to+\infty]{} F(\langle \Theta^-_{x,\mathcal{R}}(\omega), f\rangle).$$

Then, let $\mathcal{R} > 0$ and $n \in \mathbb{N}$. Since for all $z \in \mathcal{B}(x,\mathcal{R})$, $\omega(z) \leq \omega_n(z)$,

$$\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega(z))\, dz \geq \int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_n(z))\, dz.$$

Moreover, since

$$\lim_{n\to+\infty}\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz = \int_{\mathcal{B}(x,\mathcal{R})}(1-\omega(z))\,dz \geq \int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz,$$

if $\lim_{n\to+\infty}\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz = 0$, then for all $n\in\mathbb{N}$, $\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz = 0$, and thus :

$$\lim_{n\to+\infty}\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz\right) = \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega(z))\,dz\right).$$

Conversely, if $\lim_{n\to+\infty}\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz \neq 0$, since $\delta_0(\bullet)$ is continuously equal to 0 over $\mathbb{R}_+^*$,

$$\lim_{n\to+\infty}\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_n(z))\,dz\right) = \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega(z))\,dz\right).$$

We conclude by using the dominated convergence theorem. □

In order to prove Proposition 3.2.10, we will need the following result, which illustrates in which sense the ∞-parent SLFV can be considered as the limit $k\to+\infty$ of the k-parent SLFV.

Lemma 3.2.14. *Let $(M_t^\infty)_{t\geq 0}$ be the ∞-parent SLFV with initial density ω associated to Π^c. Then, for all $t\geq 0$, $(M_{k,t}^{\Pi^c,\omega})_{k\geq 2}$ converges vaguely to M_t^∞ as $k\to+\infty$.*

Proof. Let $t\geq 0$, and let ω_t^∞ be the density of the ∞-parent SLFV with initial density ω associated to Π^c, considered at time t. Let $f\in C_c(\mathbb{R}^d)$. Then f is integrable and

$$\forall x\in\mathbb{R}^d, f(x)\omega_{k,t}^{\Pi^c,\omega}(x) \xrightarrow[k\to+\infty]{} f(x)\omega_t^\infty(x)$$
$$\forall x\in\mathbb{R}^d, \left|f(x)\omega_{k,t}^{\Pi^c,\omega}(x)\right| \leq |f(x)|.$$

Therefore, by the dominated convergence theorem,

$$\lim_{k\to+\infty}\int_{\mathbb{R}^d} f(x)\omega_{k,t}^{\Pi^c,\omega}(x)dx = \int_{\mathbb{R}^d} f(x)\omega_t^\infty(x)dx$$
and $$\lim_{k\to+\infty}\int_{\mathcal{B}(x,\mathcal{R})} f(y)\omega_{k,t}^{\Pi^c,\omega}(y)dx = \int_{\mathcal{B}(x,\mathcal{R})} f(y)\omega_t^\infty(y)dy.$$

We now consider $\tilde{f}\in C_c(\mathbb{R}^d\times\{0,1\})$. Then, there exists $f_0, f_1\in C_c(\mathbb{R}^d)$ such that

$$\forall(x,\kappa)\in\mathbb{R}^d\times\{0,1\}, \tilde{f}(x,\kappa) = f_0(x)\mathbb{1}_{\{0\}}(\kappa) + f_1(x)\mathbb{1}_{\{1\}}(\kappa).$$

Therefore, for all $k\geq 2$,

$$\begin{aligned}\int_{\mathbb{R}^d\times\{0,1\}}\tilde{f}(x,\kappa)M_{k,t}^{\Pi^c,k}(dx,d\kappa) &= \int_{\mathbb{R}^d} f_0(x)\omega_{k,t}^{\Pi^c,k}(x)dx + \int_{\mathbb{R}^d} f_1(x)\left(1-\omega_{k,t}^{\Pi^c,k}(x)\right)dx\\ &\xrightarrow[k\to+\infty]{} \int_{\mathbb{R}^d} f_0(x)\omega_t^\infty(x)dx + \int_{\mathbb{R}^d} f_1(x)\left(1-\omega_t^\infty(x)\right)dx\\ &= \int_{\mathbb{R}^d\times\{0,1\}}\tilde{f}(x,\kappa)M_t^\infty(dx,d\kappa)\end{aligned}$$

and we conclude that $(M_{k,t}^{\Pi^c,\omega})_{k\geq 2}$ converges vaguely to M_t^∞ as $k\to+\infty$. □

Lemma 3.2.15. *Let $(M_t^\infty)_{t\geq 0}$ be the ∞-parent SLFV of initial condition M^0, constructed using the initial density ω and Π^c. Then, for all $F \in C^1(\mathbb{R})$ and $f \in C_c(\mathbb{R}^d)$, for all $l \geq 1$, for all $0 \leq t_1 < ... < t_l \leq t < t+s$, for all $h_1, ..., h_l \in C_b(\mathcal{M}_\lambda)$,*

$$\lim_{k\to+\infty} \mathbb{E}\left[\left(\Psi_{F,f}(M^k_{t+s}) - \Psi_{F,f}(M^k_t) - \int_t^{t+s} \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)du\right) \times \left(\prod_{i=1}^l h_i(M^k_{t_i})\right)\right] = 0.$$

Proof. For all $k \geq 2$, we set $(M^k_u)_{u\geq 0} = (M^{\Pi^c,\omega}_{k,u})_{u\geq 0}$ the k-parent SLFV associated to Π^c and with initial condition ω. Moreover, for all $u \geq 0$, let ω^k_u be a density of M^k_u.

Let $l \geq 1$, $0 \leq t_1 < ... < t_l \leq t < t+s$ and $h_1, ..., h_l \in C_b(\mathcal{M}_\lambda)$. Then, for all $k \geq 2$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\Psi_{F,f}(M^k_{t+s}) - \Psi_{F,f}(M^k_t) - \int_t^{t+s} \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)du\right) \times \left(\prod_{i=1}^l h_i(M^k_{t_i})\right)\right] \\
=&\mathbb{E}\left[\left(\Psi_{F,f}(M^k_{t+s}) - \Psi_{F,f}(M^k_t) - \int_t^{t+s} \mathcal{L}^k_\mu \Psi_{F,f}(M^k_u)du\right) \times \left(\prod_{i=1}^l h_i(M^k_{t_i})\right)\right] \\
&+ \mathbb{E}\left[\left(\int_t^{t+s} \mathcal{L}^k_\mu \Psi_{F,f}(M^k_u) - \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)du\right) \times \left(\prod_{i=1}^l h_i(M^k_{t_i})\right)\right].
\end{aligned}$$

Since $(M^k_u)_{u\geq 0}$ is solution to the martingale problem associated to $(\mathcal{L}^k, \delta_{M^k_0})$, the above is equal to

$$0 + \mathbb{E}\left[\left(\int_t^{t+s} \mathcal{L}^k_\mu \Psi_{F,f}(M^k_u) - \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)du\right) \times \left(\prod_{i=1}^l h_i(M^k_{t_i})\right)\right].$$

From Lemmas 3.5.4 and 3.5.5 in Section 3.5, we can apply the dominated convergence theorem to

$$\mathbb{E}\left[\left(\int_t^{t+s} |\mathcal{L}^k_\mu \Psi_{F,f}(M^k_u) - \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)|du\right) \times \left(\prod_{i=1}^l |h_i(M^k_{t_i})|\right)\right],$$

and we obtain

$$\begin{aligned}
&\lim_{k\to+\infty} \mathbb{E}\left[\left(\Psi_{F,f}(M^k_{t+s}) - \Psi_{F,f}(M^k_t) - \int_t^{t+s} \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)du\right) \times \left(\prod_{i=1}^l h_i(M^k_{t_i})\right)\right] \\
=&\mathbb{E}\left[\left(\int_t^{t+s} \lim_{k\to+\infty} \left(\mathcal{L}^k_\mu \Psi_{F,f}(M^k_u) - \mathcal{L}^\infty_\mu \Psi_{F,f}(M^k_u)\right) du\right) \times \left(\lim_{k\to+\infty} \prod_{i=1}^l h_i(M^k_{t_i})\right)\right],
\end{aligned}$$

assuming that the different limits exist.

Now, let $k \geq 2$ and $u \in [t, t+s]$. We have

$$
\begin{aligned}
&\mathcal{L}_\mu^k \Psi_{F,f}(M_u^k) - \mathcal{L}_\mu^\infty \Psi_{F,f}(M_u^k) \\
&= \int_0^\infty \int_{Supp^{\mathcal{R}}(f)} \Big(F(\langle \Theta_{x,\mathcal{R}}^+(\omega_u^k), f\rangle) - F(\langle \omega_u^k, f\rangle) \Big) \\
&\qquad \times \left[\int_{\mathcal{B}(x,\mathcal{R})^k} \prod_{j=1}^k \left(\frac{\omega_u^k(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_k - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) \right] \\
&\qquad + \Big(F(\langle \Theta_{x,\mathcal{R}}^-(\omega_u^k), f\rangle) - F(\langle \omega_u^k, f\rangle) \Big) \\
&\qquad \times \left[\delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) - \int_{\mathcal{B}(x,\mathcal{R})^k} \prod_{j=1}^k \left(\frac{\omega_u^k(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_k \right] dx \mu(d\mathcal{R}) \\
&= \int_0^\infty \int_{Supp^{\mathcal{R}}(f)} \Big(F(\langle \Theta_{x,\mathcal{R}}^+(\omega_u^k), f\rangle) - F(\langle \Theta_{x,\mathcal{R}}^-(\omega_u^k), f\rangle) \Big) \\
&\qquad \times \left[\int_{\mathcal{B}(x,\mathcal{R})^k} \prod_{j=1}^k \left(\frac{\omega_u^k(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_k - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) \right] dx \mu(d\mathcal{R}).
\end{aligned}
$$

The term inside the integral is bounded in absolute value, by Lemma 3.5.1 in Section 3.5. Moreover, as $(M_u^k)_{k\geq 2}$ converges vaguely to M_u^∞ by Lemma 3.2.14, we can apply Lemma 3.2.12 and we obtain

$$
\lim_{k\to+\infty} F(\langle \Theta_{x,\mathcal{R}}^+(\omega_u^k), f\rangle) - F(\langle \Theta_{x,\mathcal{R}}^-(\omega_u^k), f\rangle) = F(\langle \Theta_{x,\mathcal{R}}^+(\omega_u^\infty), f\rangle) - F(\langle \Theta_{x,\mathcal{R}}^-(\omega_u^\infty), f\rangle).
$$

Therefore, we have to show that

$$
\lim_{k\to+\infty} \int_{\mathcal{B}(x,\mathcal{R})^k} \prod_{j=1}^k \left(\frac{\omega_u^k(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_k - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) = 0.
$$

We cannot apply directly Lemma 3.2.11, because the density also depends on k. However,

$$
\begin{aligned}
&\left| \int_{\mathcal{B}(x,\mathcal{R})^k} \prod_{j=1}^k \left(\frac{\omega_u^k(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_k - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) \right| \\
&\leq \left| \int_{\mathcal{B}(x,\mathcal{R})^k} \left(\prod_{j=1}^k \frac{\omega_u^k(y_j)}{V_{\mathcal{R}}} - \prod_{j=1}^k \frac{\omega_u^\infty(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_{n_k} \right| \\
&+ \left| \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} (1 - \omega_u^\infty(y))\, dy \right) \right| \\
&+ \left| \int_{\mathcal{B}(x,\mathcal{R})^k} \prod_{j=1}^k \left(\frac{\omega_u^\infty(y_j)}{V_{\mathcal{R}}} \right) dy_1 ... dy_k - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} (1 - \omega_u^\infty(y))\, dy \right) \right|.
\end{aligned}
$$

We can apply Lemma 3.2.11 to the third term. Since for all $y \in \mathbb{R}^d$, $\omega_u^\infty(y) \leq \omega_u^k(y)$, we showed in the proof of Lemma 3.2.13 that

$$
\lim_{k\to+\infty} \left| \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} \Big(1 - \omega_u^k(y)\Big)\, dy \right) - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} (1 - \omega_u^\infty(y))\, dy \right) \right| = 0.
$$

Regarding the first term, we distinguish two cases. If $V_{\mathcal{R}}^{-1}\int_{\mathcal{B}(x,\mathcal{R})}\omega_u^{\infty}(y)dy=1$, since

$$\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{\infty}(y)}{V_{\mathcal{R}}}dy\leq\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{k}(y)}{V_{\mathcal{R}}}dy\leq 1,$$

we obtain that in fact for every $k\geq 2$

$$\int_{\mathcal{B}(x,\mathcal{R})^k}\prod_{j=1}^{k}\left(\frac{\omega_u^{k}(y_j)}{V_{\mathcal{R}}}-\prod_{j=1}^{k}\frac{\omega_u^{\infty}(y_j)}{V_{\mathcal{R}}}\right)dy_1...dy_{n_k}=0.$$

Conversely, assume $V_{\mathcal{R}}^{-1}\int_{\mathcal{B}(x,\mathcal{R})}\omega_u^{\infty}(y)dy<1$. Since,

$$\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{k}(y)}{V_{\mathcal{R}}}dy\xrightarrow[k\to+\infty]{}\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{\infty}(y)}{V_{\mathcal{R}}}dy,$$

there exist $0<M<1$ and $k'\geq 2$ such that :

$$\forall k\geq k',\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{k}(y)}{V_{\mathcal{R}}}dy\leq M.$$

Therefore,

$$\left|\int_{\mathcal{B}(x,\mathcal{R})^k}\prod_{j=1}^{k}\left(\frac{\omega_u^{k}(y_j)}{V_{\mathcal{R}}}\right)-\prod_{j=1}^{k}\left(\frac{\omega_u^{\infty}(y_j)}{V_{\mathcal{R}}}\right)dy_1...dy_k\right|=\left|\left(\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{k}(y)}{V_{\mathcal{R}}}dy\right)^k-\left(\int_{\mathcal{B}(x,\mathcal{R})}\frac{\omega_u^{\infty}(y)}{V_{\mathcal{R}}}dy\right)^k\right|$$
$$\xrightarrow[k\to+\infty]{}0,$$

and we can conclude. □

We can now show that the ∞-parent SLFV is solution of the martingale problem introduced in Proposition 3.2.10.

Proof. (Proposition 3.2.10) Let $F\in C^1(\mathbb{R})$ and $f\in C_c(\mathbb{R}^d)$. For all $k\geq 2$, we set $(M_t^k)_{t\geq 0}=(M_{k,t}^{\Pi^c,\omega})_{t\geq 0}$. Let $l\geq 1$, let $0\leq t_1<...<t_l\leq t<t+s$, and let $h_1,...,h_l\in C_b(\mathcal{M}_\lambda)$. By Lemma 3.2.15,

$$\lim_{k\to+\infty}\mathbb{E}\left[\left(\Psi_{F,f}(M_{t+s}^k)-\Psi_{F,f}(M_t^k)-\int_t^{t+s}\mathcal{L}_\mu^{\infty}\Psi_{F,f}(M_u^k)du\right)\times\left(\prod_{i=1}^{l}h_i(M_{t_i}^k)\right)\right]=0.$$

Since $(M_{t+s}^k)_{k\geq 2}$ (resp. $(M_t^k)_{k\geq 2}$) converges vaguely to M_{t+s}^{∞} (resp. M_t^{∞}) by Lemma 3.2.14, we can apply Lemma 3.2.12 and we obtain

$$\lim_{k\to+\infty}\Psi_{F,f}(M_{t+s}^k)=\Psi_{F,f}(M_{t+s}^{\infty})$$
$$\text{and }\lim_{k\to+\infty}\Psi_{F,f}(M_t^k)=\Psi_{F,f}(M_t^{\infty}).$$

Moreover, by Lemma 3.2.13, for all $u\in[t,t+s]$,

$$\lim_{k\to+\infty}\mathcal{L}_\mu^{\infty}\Psi_{F,f}(M_u^k)=\mathcal{L}_\mu^{\infty}\Psi_{F,f}(M_u^{\infty}),$$

which is uniformly bounded in $M\in\mathcal{M}_\lambda$ by Lemma 3.5.5 in Section 3.5. Since for all $i\in[\![1,l]\!]$, $h_i\in C_b(\mathcal{M}_\lambda)$,

$$\forall i\in[\![1,l]\!],\lim_{k\to+\infty}h_i(M_{t_i}^k)=h_i(M_{t_i}^{\infty}).$$

Therefore, by Eq.(3.2.2) and by Lemmas 3.5.4, 3.5.5 in Section 3.5, we can apply the dominated convergence theorem and obtain

$$\mathbb{E}\left[\left(\Psi_{F,f}(M^{\infty}_{t+s}) - \Psi_{F,f}(M^{\infty}_{t}) - \int_t^{t+s} \mathcal{L}^{\infty}_{\mu}\Psi_{F,f}(M^{\infty}_{u})du\right) \times \left(\prod_{i=1}^{l} h_i(M^{\infty}_{t_i})\right)\right] = 0.$$

We conclude that

$$\left(\Psi_{F,f}(M^{\infty}_{t}) - \Psi_{F,f}(M^{\infty}_{0}) - \int_0^{t} \mathcal{L}^{\infty}_{\mu}\Psi_{F,f}(M^{\infty}_{u})du\right)_{t\geq 0}$$

is indeed a martingale. □

3.3 The ∞-parent ancestral process : definition and characterization

3.3.1 Definition and first properties

In order to show that the ∞-parent ancestral process $(\Xi^{\infty}_t)_{t\geq 0}$ introduced in Definition 3.1.11 is well-defined, we start by observing that the only reproduction events affecting Ξ^{∞}_t are the ones intersecting its boundary $\overline{\Xi^{\infty}_t}\backslash\mathring{\Xi}^{\infty}_t$. Therefore, it is sufficient to consider only the reproduction events affecting its border, or the ones affecting a well-chosen space containing it.

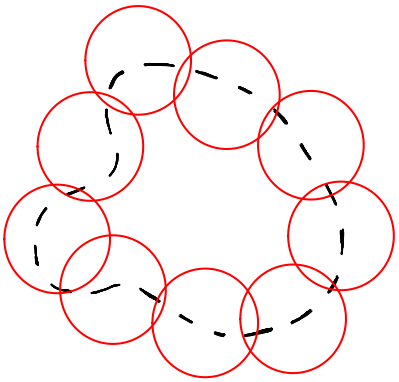
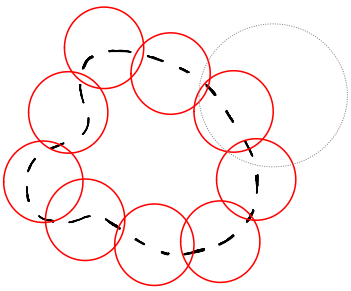


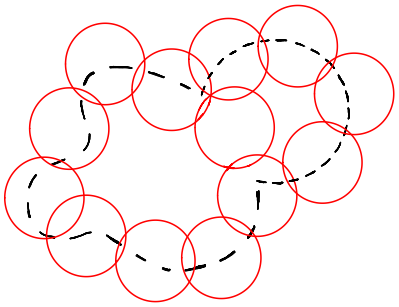
Figure 3.1: Initial state of the ∞-parent ancestral process (dashed line), and a covering of its border by balls of radius $\widetilde{\mathcal{R}}$.

In order to control the rate at which the ∞-parent ancestral process jumps, we start by taking $\widetilde{\mathcal{R}} > 0$ satisfying some condition which will be introduced later, and we cover the border $\overline{\Xi^{\infty}_0}\backslash\mathring{\Xi}^{\infty}_0$ of Ξ^{∞}_0 with balls of radius $\widetilde{\mathcal{R}}$ (see Figure 3.1). Then, informally, whenever a reproduction event overlaps what we will call the $\widetilde{\mathcal{R}}$-covering :

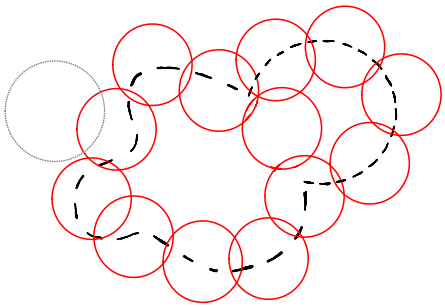
- if this reproduction event has a radius of at most $\widetilde{\mathcal{R}}$, it is included in the ball of same center but of radius $\widetilde{\mathcal{R}}$. We add this ball of radius $\widetilde{\mathcal{R}}$ to the covering.
- Otherwise, we cover the border of the area of the reproduction event by balls of radius $\widetilde{\mathcal{R}}$, and we add these balls to the covering.



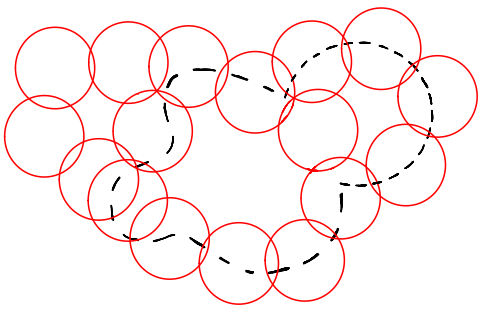
(a) Reproduction event (grey line) affecting the ∞-parent ancestral process at time $t > 0$.



(b) The ∞-parent ancestral process is updated, and a covering of the border of the reproduction event by balls of radius $\widetilde{\mathcal{R}}$ is added to the $\widetilde{\mathcal{R}}$-covering process.



(c) Since the $\widetilde{\mathcal{R}}$-covering process is bigger than the border of the ∞-parent ancestral process, it can be affected by reproduction events (grey line) which do not intersect the ∞-parent ancestral process.



(d) Updated $\widetilde{\mathcal{R}}$-covering process after a reproduction event affecting it while not intersecting the ∞-parent ancestral process.

Figure 3.2: Illustration of the dynamics of the ∞-parent ancestral process (dashed line) and its associated $\widetilde{\mathcal{R}}$-covering process.

See Figure 3.2 for an illustration of this dynamics.

Note that since the covering contains the border $\overline{\Xi_t^\infty} \backslash \mathring{\Xi}_t^\infty$ of Ξ_t^∞ but is not equal to it, there are more reproduction events affecting the $\widetilde{\mathcal{R}}$-covering than reproduction events affecting Ξ_t^∞.

Constructed this way, the $\widetilde{\mathcal{R}}$-covering contains only balls of radius $\widetilde{\mathcal{R}}$, each one being overlapped by a reproduction event at rate

$$\int_0^\infty V_1(\widetilde{\mathcal{R}} + r)^d \mu(dr).$$

Moreover, since the covering is constructed using the same Poisson point process as for $(\Xi_t^\infty)_{t\geq 0}$, at any time t the current state of the covering contains the border $\overline{\Xi_t^\infty} \backslash \mathring{\Xi}_t^\infty$ of Ξ_t^∞. Since the rate at which $(\Xi_t^\infty)_{t\geq 0}$ jumps is bounded by the rate at which the covering we just constructed is updated, we can show that $(\Xi_t^\infty)_{t\geq 0}$ is well-defined by controlling the rate at which new balls are added to the $\widetilde{\mathcal{R}}$-covering.

Let us now define the border covering process we just introduced rigorously.

Definition 3.3.1 (Border covering process)**.** *In the notation of Definition 3.1.11, let $\widetilde{\mathcal{R}} > 0$ be such that μ satisfies Condition (3.1.5). Let $x_1, ..., x_N \in \mathbb{R}^d$, $N \geq 1$ be such that initially the border of Ξ^0 is entirely covered by the N balls of radius $\widetilde{\mathcal{R}}$ $(B(x_i, \widetilde{\mathcal{R}}))_{1 \leq i \leq N}$. Then, the $\widetilde{\mathcal{R}}$-covering process $(C_t)_{t \geq 0}$ associated to $(\Xi_t^\infty)_{t \geq 0}$ is constructed in the following way.*

Let $\widetilde{\Pi}$ be a Poisson point process on $\mathbb{R} \times \mathbb{R}^d \times (0, +\infty)$ with intensity $dt \otimes dx \otimes \mu(dr)$. First, we set $C_0 = \{x_1, ..., x_N : 1 \leq i \leq N\}$. Then, for all $(t, x, \mathcal{R}) \in \widetilde{\Pi}$, if $C_{t-} \cap \mathcal{B}(x, \mathcal{R}) \neq \emptyset$, let $n \in \mathbb{N}^*$ such that $(n-1)\widetilde{\mathcal{R}} \leq \mathcal{R} \leq n\widetilde{\mathcal{R}}$. We construct a covering of the border of $\mathcal{B}(x, \mathcal{R})$ by at most $a_d \times n^{d-1}$ balls of radius $\widetilde{\mathcal{R}}$, and C_t is obtained by adding the center of these balls to C_{t-}.

The interest of the border covering process lies in the fact that, as we argued earlier, for all $t \geq 0$,

$$\overline{\Xi_t^\infty} \backslash \mathring{\Xi_t^\infty} \subseteq \bigcup_{x \in C_t} \mathcal{B}(x, \widetilde{\mathcal{R}}).$$

Therefore, the jump rate of Ξ_t^∞ is bounded above by

$$Card(C_t) \times \int_0^\infty V_1(\widetilde{\mathcal{R}} + r)^d \mu(dr).$$

Lemma 3.3.2. *In the notation of Definitions 3.1.11 and 3.3.1, $(Card(C_t))_{t \geq 0}$ is bounded from above by $(Y_t)_{t \geq 0}$ the number of particles in a branching process in which each particle branches independently of the others at rate*

$$\int_0^\infty V_1(\widetilde{\mathcal{R}} + r)^d \mu(dr),$$

and in which at each branching event, the number of descendants is equal to $a_d \times n^{d-1} + 1$, $n \geq 1$ with probability

$$\frac{\int_{(n-1)\widetilde{\mathcal{R}}}^{n\widetilde{\mathcal{R}}} (\widetilde{\mathcal{R}} + r)^d \mu(dr)}{\int_0^\infty (\widetilde{\mathcal{R}} + r)^d \mu(dr)}.$$

Moreover, for all $t \geq 0$, $Y_t < +\infty$ p.s, and $\mathbb{E}[Y_t] < +\infty$.

Proof. How to construct the branching process $(Y_t)_{t \geq 0}$ from $(C_t)_{t \geq 0}$ is clear. The jump rates and transition probabilities come from the fact that for any point $x \in C_t$ and for all $n \geq 1$, the ball $\mathcal{B}(x, \widetilde{\mathcal{R}})$ is affected by a reproduction event of radius $(n-1)\widetilde{\mathcal{R}} \leq \mathcal{R} \leq n\widetilde{\mathcal{R}}$ at rate

$$\int_{(n-1)\widetilde{\mathcal{R}}}^{n\widetilde{\mathcal{R}}} V_1(\widetilde{\mathcal{R}} + \mathcal{R})^d \mu(d\mathcal{R}),$$

and such a reproduction event generates $a_d \times n^{d-1}$ new balls in the border covering process.

Then, if Φ is the probability generating function of the number of descendants,

$$\Phi'(1) = \sum_{n=1}^{+\infty} \left(\int_{(n-1)\widetilde{\mathcal{R}}}^{n\widetilde{\mathcal{R}}} V_1(\widetilde{\mathcal{R}} + r)^d \mu(dr) \right) \times (a_d \times n^{d-1} + 1) < +\infty$$

since μ satisfies Condition (3.1.5). Therefore, by Theorem III.2.1 in [AN72], Y_t is finite for all $t \geq 0$ a.s, and $\mathbb{E}[Y_t] < +\infty$ for all $t \geq 0$. □

We can then conclude that $(\Xi_t^\infty)_{t \geq 0}$ is well-defined and Markovian using the fact that the jump rate of Ξ_t^∞ is bounded from above by

$$Y_t \times \int_0^\infty V_1 \times (\widetilde{\mathcal{R}} + \mathcal{R})^d \mu(d\mathcal{R}) < +\infty \text{ p.s },$$

and proceeding as in the proof of Proposition 1.5 from [EVY20].

3.3.2 Characterization via a martingale problem

The goal of this section is to introduce how to characterize the ∞-parent ancestral process as the unique solution to a martingale problem.

In all that follows, let $F \in C_b^1(\mathbb{R})$ and $f \in \mathbb{B}(\mathbb{R}^d)$. We extend the definition of the function $\Phi_{F,f}$ to the space of measures $m(E) \in \mathcal{M}^{cf}$, setting

$$\begin{aligned}\Phi_{F,f}(m(E)) :=& F\left(\int_{\mathbb{R}^d} f(x)m(E)dx\right)\\ =& F\left(\int_E f(x)dx\right).\end{aligned}$$

Moreover, for all $E \in \mathcal{E}^{cf}$ and $\mathcal{R} > 0$, we set

$$S^{\mathcal{R}}(E) := \{x \in \mathbb{R}^d : \exists y \in E, ||x-y|| \leq \mathcal{R}\}.$$

Note that this definition is reminiscent of the definition of $S^{\mathcal{R}}(\Xi)$ with $\Xi \in \mathcal{M}_p(\mathbb{R}^d)$.

Let μ be a σ-finite measure on $\mathbb{R}_+^*$ satisfying Condition (3.1.5). We define the operator $\mathcal{G}_\mu^\infty$ on functions of the form $\Phi_{F,f}$ the following way. For all $m(E) \in \mathcal{M}^{cf}$, we set

$$\mathcal{G}_\mu^\infty \Phi_{F,f}(m(E)) := \int_0^\infty \int_{S^{\mathcal{R}}(E)} F\left(\langle m(E \cup \mathcal{B}(x,\mathcal{R})), f\rangle\right) - F\left(\langle m(E), f\rangle\right) dx\mu(d\mathcal{R}).$$

We show in Section 3.5 that this operator is well-defined, and give some properties that it satisfies. The ∞-parent ancestral process is then solution to the following martingale problem.

Proposition 3.3.3. *Let μ be a σ-finite measure on $(0,+\infty)$ satisfying Condition (3.1.5). Let $\Xi^0 \in \mathcal{M}^{cf}$, and let $(\Xi_t^\infty)_{t\geq 0}$ be the ∞-parent ancestral process associated to μ with initial condition Ξ^0.*

Then, for all $F \in C_b^1(\mathbb{R})$ and for all measurable function $f : \mathbb{R}^d \to \{0,1\}$, the process

$$\left(\Phi_{F,f}(\Xi_t^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_s^\infty)ds\right)_{t\geq 0}$$

is a martingale.

Proof. Let $(\mathcal{F}_t)_{t\geq 0}$ be the filtration generated by $(\Xi_t)_{t\geq 0}$, and let $0 \leq s \leq t$.

By Lemma 3.5.7 in Section 3.5,

$$\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right]$$

is well-defined, and

$$\begin{aligned}&\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right]\\ =&\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty) - \Phi_{F,f}(\Xi_s^\infty) - \int_s^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right]\\ &+\mathbf{E}\left[\Phi_{F,f}(\Xi_s^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^s \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right]\\ =&\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty)|\,\Xi_s^\infty\right] - \Phi_{F,f}(\Xi_s^\infty) - \mathbf{E}\left[\int_s^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \Xi_s^\infty\right] + \Phi_{F,f}(\Xi_s^\infty) - \Phi_{F,f}(\Xi_0^\infty)\\ &-\int_0^s \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du\end{aligned}$$

since $(\Xi_u^\infty)_{u\geq 0}$ is Markovian. Let $(\widetilde{\Xi}_u)_{u\geq 0}$ be another ∞-parent ancestral process associated to μ, this time with initial condition Ξ_s^∞. Then,

$$\begin{aligned}
&\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right] \\
=&\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty)|\,\Xi_s^\infty\right] - \Phi_{F,f}(\Xi_s^\infty) - \mathbf{E}\left[\int_0^{t-s} \mathcal{G}_\mu^\infty \Phi_{F,f}(\widetilde{\Xi}_u)du \middle| \Xi_s^\infty\right] + \Phi_{F,f}(\Xi_s^\infty) - \Phi_{F,f}(\Xi_0^\infty) \\
&- \int_0^s \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du.
\end{aligned}$$

By Lemmas 3.5.7 and 3.5.8 in Section 3.5,

$$\begin{aligned}
\mathbf{E}\left[\int_0^{t-s} \mathcal{G}_\mu^\infty \Phi_{F,f}(\widetilde{\Xi}_u)du \middle| \Xi_s^\infty\right] &= \int_0^{t-s} \mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F,f}(\widetilde{\Xi}_u) \middle| \Xi_s^\infty\right] du \\
&= \int_0^{t-s} \frac{d}{dv}\mathbf{E}\left[\Phi_{F,f}(\widetilde{\Xi}_v) \middle| \Xi_s^\infty\right]\Big|_{v=u} du \\
&= \mathbf{E}\left[\Phi_{F,f}\left(\widetilde{\Xi}_{t-s}\right)|\Xi_s^\infty\right] - \mathbf{E}\left[\Phi_{F,f}\left(\widetilde{\Xi}_0\right)|\Xi_s^\infty\right] \\
&= \mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty)|\Xi_s^\infty\right] - \Phi_{F,f}(\Xi_s^\infty).
\end{aligned}$$

Therefore, we obtain

$$\mathbf{E}\left[\Phi_{F,f}(\Xi_t^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right] = \Phi_{F,f}(\Xi_s^\infty) - \Phi_{F,f}(\Xi_0^\infty) - \int_0^s \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_u^\infty)du$$

and we can conclude. □

3.4 Uniqueness of the solution to the martingale problem characterizing the ∞-parent SLFV

In order to show the uniqueness of the solution to the martingale problem characterizing the ∞-parent SLFV, we first need to extend the set of functions over which the operators $\mathcal{L}_\mu^\infty$ and $\mathcal{G}_\mu^\infty$ are defined.

3.4.1 Extended martingale problem for the ∞-parent SLFV

For all $\alpha \in \mathbb{R}$, we set $F^\alpha : x \to \delta_\alpha(x)$, and for all $E \in \mathcal{E}^{cf}$, we set $f^E : x \to \mathbb{1}_{x\in E}$. Let μ be a σ-finite measure on $(0,+\infty)$ satisfying Condition (3.1.4), and let $M^0 \in \mathcal{M}_\lambda$. The goal of this section is to prove the following result.

Lemma 3.4.1. *Let M be a solution to the martingale problem associated to $(\mathcal{L}_\mu^\infty, \delta_{M^0})$. Then, for all $E \in \mathcal{E}^{cf}$,*

$$\left(\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_t) - \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_0) - \int_0^t \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_s)ds\right)_{t\geq 0}$$

is a martingale, where $\Psi_{F^{\mathrm{Vol}(E)},f^E} : M \in \mathcal{M}_\lambda \to \Psi_{F^{\mathrm{Vol}(E)},f^E}(M)$ is the function defined by

$$\begin{aligned}
\forall M \in \mathcal{M}_\lambda, \Psi_{F^{\mathrm{Vol}(E)},f^E}(M) :&= F^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right) \\
&= \delta_{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right) \\
&= \delta_0\left(\mathrm{Vol}(E) - \langle \omega_M, f^E\rangle\right).
\end{aligned}$$

This lemma is a direct consequence of the following lemma.

Lemma 3.4.2. *Let M be a solution to the martingale problem associated to $(\mathcal{L}_\mu^\infty, \delta_{M^0})$. Then, for all $E \in \mathcal{E}^{cf}$, for all $l \geq 1$, for all $0 \leq t_1 < ... < t_l \leq t < t+s$, for all $h_1, ..., h_l \in C_b(\mathcal{M}_\lambda)$,*

$$\mathbb{E}\left[\left(\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_{t+s}) - \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_t) - \int_t^{t+s} \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_u)du\right) \times \left(\prod_{i=1}^{l} h_i(M_{t_i})\right)\right] = 0$$

Let $E \in \mathcal{E}^{cf}$. Let $(F_n^{\mathrm{Vol}(E)})_{n\in\mathbb{N}} \in C^1(\mathbb{R})$ and $(f_n^E)_{n\in\mathbb{N}} \in C_c(\mathbb{R}^d)$ be two sequences satisfying the following conditions.

(A) $F_n^{\mathrm{Vol}(E)} \xrightarrow[n\to+\infty]{} F^{\mathrm{Vol}(E)}$ pointwise and in L^1,
(B) $f_n^E \xrightarrow[n\to+\infty]{} f^E$ pointwise and in L^1,
(C) $\forall n \in \mathbb{N}, \forall x \in \mathbb{R}, 0 \leq F_n^{\mathrm{Vol}(E)}(x) \leq 1$ and $F_n^{\mathrm{Vol}(E)}(\mathrm{Vol}(E)) = 1$,
(D) $\forall n \in \mathbb{N}, \forall x \in \mathbb{R}^d, 0 \leq f_n^E(x) \leq 1$ and $\forall z \in E, f_n^E(z) = 1$,
(E) $\forall n \in \mathbb{N}, F_n^{\mathrm{Vol}(E)}$ is increasing over $(-\infty, \mathrm{Vol}(E)]$ and decreasing over $[\mathrm{Vol}(E), +\infty)$,
(F) $\forall n \in \mathbb{N}, \mathrm{Vol}(Supp(f_n^E)\backslash E) \leq n^{-1}$, and $Supp(f_{n+1}^E) \subseteq Supp(f_n^E)$
(G) $\forall n \in \mathbb{N}, F_n^{\mathrm{Vol}(E)}\left(\mathrm{Vol}(E) + n^{-1}\right) \geq 1 - n^{-1}$ and $F_n^{\mathrm{Vol}(E)}\left(\mathrm{Vol}(E) - n^{-1}\right) \geq 1 - n^{-1}$.

First, we observe that since $F^{\mathrm{Vol}(E)}$ and $(F_n^{\mathrm{Vol}(E)})_{n\in\mathbb{N}^*}$ are bounded by one (by Hypothesis (C)), for all $M \in \mathcal{M}_\lambda$ and $n \in \mathbb{N}^*$

$$\left|\Psi_{F^{\mathrm{Vol}(E)},f^E}(M)\right| \leq 1 \tag{3.4.1}$$

$$\left|\Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M)\right| \leq 1 \tag{3.4.2}$$

$$\left|\Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M)\right| \leq 1 \tag{3.4.3}$$

Moreover, there exists $C_E > 0$ such that for all $\mathcal{R} > 0$,

$$\mathrm{Vol}(S^{\mathcal{R}}(E)) \leq C^E \times \left(\mathcal{R}^d \vee 1\right), \tag{3.4.4}$$

where we recall that $S^{\mathcal{R}}(E)$ is defined by

$$S^{\mathcal{R}}(E) := \{x \in \mathbb{R}^d : \exists y \in E, ||x - y|| \leq \mathcal{R}\}.$$

Therefore, we have the following lemma.

Lemma 3.4.3. *There exists $C_2^E > 0$ such that for all $M \in \mathcal{M}_\lambda$ and $n \in \mathbb{N}^*$,*

$$\left|\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M)\right| \leq C_2^E$$
$$\left|\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M)\right| \leq C_2^E$$
$$\left|\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M)\right| \leq C_2^E.$$

Proof. Let $M \in \mathcal{M}_\lambda$.

$$\begin{aligned}\left|\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M)\right| \leq & \int_0^\infty \int_{S^{\mathcal{R}}(E)} \left|1-\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_M(z))\,dz\right)\right| \\ & \times \left|F^{\mathrm{Vol}(E)}\left(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f^E\rangle\right) - F^{\mathrm{Vol(E)}}\left(\langle \omega_M, f^E\rangle\right)\right| dx\mu(d\mathcal{R}) \\ \leq & \int_0^\infty \int_{S^{\mathcal{R}}(E)} 2dx\mu(d\mathcal{R}) \\ \leq & 2 \times \int_0^\infty C^E \times \left(\mathcal{R}^d \vee 1\right)\mu(d\mathcal{R}) \\ < & +\infty\end{aligned}$$

since μ satisfies Condition (3.1.4). Here we passed from line 1 to line 2 using the fact that $F^{\mathrm{Vol}(E)}$ is bounded by 1, and from line 2 to line 3 using Eq. (3.4.4).

Setting $C_2^E = 2C^E \times \int_0^\infty \left(\mathcal{R}^d \vee 1\right)\mu(d\mathcal{R})$, we obtain

$$\left|\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M)\right| \leq C_2^E.$$

Similarly, we can show that for all $n \in \mathbb{N}^*$,

$$\begin{aligned} & \left|\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M)\right| \leq C_2^E \\ \text{and} \quad & \left|\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M)\right| \leq C_2^E. \end{aligned}$$

□

This lemma, along with Eqs. (3.4.1, 3.4.2, 3.4.3, 3.4.4), will allow us to use the dominated convergence theorem in the proof of Lemma 3.4.2.

Since by Hypothesis (A) the sequence $(F_n^{\mathrm{Vol}(E)})_{n\in\mathbb{N}^*}$ converges pointwise to $F^{\mathrm{Vol}(E)}$, we obtain that

$$\forall M \in \mathcal{M}_\lambda, \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) \xrightarrow[n\to+\infty]{} \Psi_{F^{\mathrm{Vol}(E)},f^E}(M). \tag{3.4.5}$$

We want to show a similar result regarding $\left(\Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M)\right)_{n\in\mathbb{N}^*}$.

Lemma 3.4.4. *For all $M \in \mathcal{M}_\lambda$,*

$$\Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M) - \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) \xrightarrow[n\to+\infty]{} 0.$$

Proof. Let $M \in \mathcal{M}_\lambda$. We distinguish two cases.

Case 1 : $\int_E \omega_M(z)dz = \mathrm{Vol}(E)$.

Let $n \in \mathbb{N}^*$. Then, since by Hypothesis (D) we have $E \subseteq Supp(f_n^E)$,

$$\begin{aligned}\mathrm{Vol}(E) \leq \langle \omega_M, f_n^E\rangle & \leq \mathrm{Vol}(E) + \int_{Supp(f_n^E)\setminus E} f_n^E(z)\omega_M(z)dz \\ & \leq \mathrm{Vol}(E) + \mathrm{Vol}(Supp(f_n^E)\setminus E) \\ & \leq \mathrm{Vol}(E) + \frac{1}{n}\end{aligned}$$

using Hypotheses (D) and (F). Therefore, since $F_n^{\mathrm{Vol}(E)}$ is decreasing over $[\mathrm{Vol}(E), +\infty)$ by Hypothesis (E),

$$F_n^{\mathrm{Vol}(E)}(\mathrm{Vol}(E)) \geq \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M) \geq F_n^{\mathrm{Vol}(E)}(\mathrm{Vol}(E) + \frac{1}{n})$$

or, in other words,

$$1 \geq \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M) \geq 1 - \frac{1}{n}$$

by Hypothesis (C) and (G). Moreover,

$$\begin{aligned}\Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) &= F_n^{\mathrm{Vol}(E)}\left(\int_E \omega_M(z)dz\right)\\ &= F_n^{\mathrm{Vol}(E)}(\mathrm{Vol}(E))\\ &= 1\end{aligned}$$

by Hypothesis (C), and we can conclude.

Case 2 : $\int_E \omega_M(z)dz < \mathrm{Vol}(E)$.

Let $N \in \mathbb{N}^*$ be such that $N^{-1} \leq 2^{-1} \times \left(\mathrm{Vol}(E) - \int_E \omega_M(z)dz\right)$. Then, for all $n \geq N$, using Hypotheses (D) and (F),

$$\begin{aligned}0 \leq \langle \omega_M, f_n^E\rangle &\leq \int_E \omega_M(z)dz + \int_{Supp(f_n^E)\backslash E} \omega_M(z)dz\\ &\leq \int_E \omega_M(z)dz + \mathrm{Vol}(Supp(f_n^E)\backslash E)\\ &\leq \int_E \omega_M(z)dz + \frac{1}{n}\\ &\leq \int_E \omega_M(z)dz + \frac{1}{N}\\ &\leq \frac{1}{2} \times \int_E \omega_M(z)dz + \frac{1}{2} \times \mathrm{Vol}(E)\\ &< \mathrm{Vol}(E),\end{aligned}$$

so by Hypothesis (E),

$$\Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M) \xrightarrow[n\to+\infty]{} 0.$$

Moreover, since $\langle \omega_M, f^E\rangle < \mathrm{Vol}(E)$, again by Hypothesis (E),

$$\Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) \xrightarrow[n\to+\infty]{} 0,$$

and we can conclude. □

We now prove a similar result involving $\mathcal{L}_\mu^\infty$.

Lemma 3.4.5. *For all $M \in \mathcal{M}_\lambda$,*

$$\begin{aligned}\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) - \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M) &\xrightarrow[n\to+\infty]{} 0,\\ \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) - \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M) &\xrightarrow[n\to+\infty]{} 0.\end{aligned}$$

Proof. Let $M \in \mathcal{M}_\lambda$, and let $n \in \mathbb{N}^*$. We have

$$\begin{aligned}
&\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) - \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M) \\
&= \int_0^\infty \int_{S^{\mathcal{R}}(E)} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_M(z))\,dz\right)\right) \\
&\qquad \times \left(F_n^{\mathrm{Vol}(E)}\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f^E\rangle\right) - F_n^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right)\right) dx\mu(d\mathcal{R}) \\
&- \int_0^\infty \int_{S^{\mathcal{R}}(E)} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_M(z))\,dz\right)\right) \\
&\qquad \times \left(F^{\mathrm{Vol}(E)}\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f^E\rangle\right) - F^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right)\right) dx\mu(d\mathcal{R}) \\
&= \int_0^\infty \int_{S^{\mathcal{R}}(E)} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_M(z))\,dz\right)\right) \\
&\qquad \times \left(F_n^{\mathrm{Vol}(E)}\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f^E\rangle\right) - F^{\mathrm{Vol}(E)}\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f^E\rangle\right)\right) dx\mu(d\mathcal{R}) \\
&+ \int_0^\infty \int_{S^{\mathcal{R}}(E)} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_M(z))\,dz\right)\right) \\
&\qquad \times \left(F^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right) - F_n^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right)\right) dx\mu(d\mathcal{R}).
\end{aligned}$$

By Eq. (3.4.5), for all $x \in \mathbb{R}^d$ and $\mathcal{R} > 0$,

$$\begin{aligned}
F_n^{\mathrm{Vol}(E)}\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f^E\rangle\right) - F^{\mathrm{Vol}(E)}\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f^E\rangle\right) &\xrightarrow[n\to+\infty]{} 0 \\
\text{and} \qquad F^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right) - F_n^{\mathrm{Vol}(E)}\left(\langle \omega_M, f^E\rangle\right) &\xrightarrow[n\to+\infty]{} 0.
\end{aligned}$$

Therefore, using the bounds from the proof of Lemma 3.4.3, we can apply the dominated convergence theorem and obtain

$$\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) - \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M) \xrightarrow[n\to+\infty]{} 0.$$

We can similarly show that

$$\mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M) - \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M) \xrightarrow[n\to+\infty]{} 0$$

using Lemma 3.4.4 instead of Eq. (3.4.5). □

We can now prove Lemma 3.4.2, from which we will directly deduce Lemma 3.4.1.

Proof. (Lemma 3.4.2) Let $l \geq 1$, let $0 \leq t_1 < ... < t_l \leq t < t+s$ and let $h_1, ..., h_l \in C_b(\mathcal{M}_\lambda)$. Let $n \in \mathbb{N}^*$. Then,

$$\begin{aligned}
\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_{t+s}) =& \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_{t+s}) - \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_{t+s}) \\
&+ \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_{t+s}) - \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_{t+s}) + \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_{t+s}) \\
\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_t) =& \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_t) - \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_t) \\
&+ \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_t) - \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_t) + \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_t),
\end{aligned}$$

and for all $u \in [t, t+s]$,

$$\begin{aligned}\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_u) =& \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_u) - \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_u) \\ &+ \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_u) - \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_u) + \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_u).\end{aligned}$$

Since M is a solution of the martingale problem associated to $(\mathcal{L}_\mu^\infty, \delta_{M^0})$, for all $n \in \mathbb{N}^*$,

$$\mathbb{E}\left[\left(\Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_{t+s}) - \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_t) - \int_t^{t+s} \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_u)du\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] = 0.$$

Therefore, since all the equations written above are true for all $n \in \mathbb{N}^*$,

$$\begin{aligned}&\mathbb{E}\left[\left(\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_{t+s}) - \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_t) - \int_t^{t+s} \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_u)du\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] \\ &= \lim_{n\to+\infty} \mathbb{E}\left[\left(\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_{t+s}) - \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_{t+s})\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] \\ &\quad + \lim_{n\to+\infty} \mathbb{E}\left[\left(\Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_{t+s}) - \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_{t+s})\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] \\ &\quad - \lim_{n\to+\infty} \mathbb{E}\left[\left(\Psi_{F^{\mathrm{Vol}(E)},f^E}(M_t) - \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_t)\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] \\ &\quad - \lim_{n\to+\infty} \mathbb{E}\left[\left(\Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_t) - \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_t)\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] \\ &\quad - \lim_{n\to+\infty} \mathbb{E}\left[\left(\int_t^{t+s} \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E)},f^E}(M_u) - \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_u)du\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right] \\ &\quad - \lim_{n\to+\infty} \mathbb{E}\left[\left(\int_t^{t+s} \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f^E}(M_u) - \mathcal{L}_\mu^\infty \Psi_{F_n^{\mathrm{Vol}(E)},f_n^E}(M_u)du\right) \times \left(\prod_{i=1}^l h_i(M_{t_i})\right)\right]\end{aligned}$$

under the condition that all these limits exist.

By Eq. (3.4.5), Lemma 3.4.4 and Lemma 3.4.5, all the terms inside the expectations converge to 0 when $n \to +\infty$. Using the bounds given by Eq. (3.4.1), (3.4.2), (3.4.3) and Lemma 3.4.3, we can apply the dominated convergence theorem and obtain the desired result.

□

3.4.2 Extended martingale problem for the ∞-parent ancestral process

In this section, we prove the following result.

Lemma 3.4.6. *Let μ be a σ-finite measure on $(0, +\infty)$ satisfying Condition (3.1.5). Let $\Xi^0 \in \mathcal{M}^{cf}$, and let $(\Xi_t^\infty)_{t\geq 0} = (m(E_t))_{t\geq 0}$ be the ∞-parent ancestral process associated to μ with initial condition Ξ^0.*

Then, for all measurable function $f : \mathbb{R}^d \to \{0, 1\}$,

$$\left(\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_s^\infty)ds\right)_{t\geq 0}$$

is a martingale, where $\Phi_{\delta_0,f} : \Xi \in \mathcal{M}^{cf} \to \Phi_{\delta_0,f}(\Xi)$ *is the function defined by*

$$\forall m(E) \in \mathcal{M}^{cf}, \Phi_{\delta_0,f}(m(E)) := \delta_0 \left(\int_E f(x)dx \right).$$

Proof. Let $(\mathcal{F}_t)_{t\geq 0}$ be the filtration generated by $(\Xi_t^\infty)_{t\geq 0}$. Let $(F_n)_{n\in\mathbb{N}^*} \in C_b^1(\mathbb{R})$ be a sequence of functions converging pointwise to δ_0 such that

(A) $\forall n \in \mathbb{N}^*, F_n$ is increasing on $\mathbb{R}_-$ and decreasing on $\mathbb{R}_+$,
(B) $\forall n \in \mathbb{N}^*, F_n(0) = 1$ and $\forall x \in \mathbb{R}, 0 \leq F_n(x) \leq 1$,
(C) $\forall n \in \mathbb{N}^*, Supp(F_n) \subseteq [-n^{-3}, n^{-3}]$.

The interest of this sequence lies in the fact that for all $n \in \mathbb{N}^*$ and for all measurable function $f : \mathbb{R}^d \to \{0, 1\}$,

$$\left(\Phi_{F_n,f}(\Xi_t^\infty) - \Phi_{F_n,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_s^\infty)ds \right)_{t\geq 0}$$

is a martingale.

Let $f : \mathbb{R}^d \to \{0, 1\}$ be a measurable function, and let $0 \leq s \leq t$. $\Phi_{\delta_0,f}$ is bounded by 1, and by Hypothesis (B), the functions $(\Phi_{F_n,f})_{n\in\mathbb{N}^*}$ are bounded by 1 as well. Moreover, since $u \to \mathrm{Vol}(\Xi_u^\infty))$ is increasing, and as there exists $C_t > 0$ such that for all $\mathcal{R} > 0$,

$$\mathrm{Vol}(S^{\mathcal{R}}(\Xi_t^\infty)) \leq C_t \times \left(\mathcal{R}^d \vee 1 \right),$$

we can deduce that for all $s \in [0, t]$ and for all $n \in \mathbb{N}^*$, by Hypothesis (B),

$$\begin{aligned} \left| \mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_s^\infty) \right| &\leq \int_0^\infty 2 \times \mathrm{Vol}(S^{\mathcal{R}}(\Xi_s^\infty))\mu(d\mathcal{R}) \\ &\leq \int_0^\infty 2 \times \mathrm{Vol}(S^{\mathcal{R}}(\Xi_t))\mu(d\mathcal{R}) \\ &\leq 2C_t \times \int_0^\infty \left(\mathcal{R}^d \vee 1 \right) \mu(d\mathcal{R}). \end{aligned}$$

Similarly, we obtain that

$$\left| \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_s^\infty) \right| \leq 2C_t \times \int_0^\infty \left(\mathcal{R}^d \vee 1 \right) \mu(d\mathcal{R}).$$

Since μ satisfies Condition (3.1.4), both quantities are finite. Therefore, by Fubini's theorem, for all $n \in \mathbb{N}^*$,

$$\begin{aligned} &\mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s \right] \\ =&\mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{F_n,f}(\Xi_t^\infty) \middle| \mathcal{F}_s \right] + \mathbf{E}\left[\Phi_{F_n,f}(\Xi_0^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty) \middle| \mathcal{F}_s \right] \\ &+ \int_0^t \mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty) - \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty) \middle| \mathcal{F}_s \right] du \\ &+ \mathbf{E}\left[\Phi_{F_n,f}(\Xi_t^\infty) - \Phi_{F_n,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s \right]. \end{aligned}$$

Using Proposition 3.3.3, we obtain that

$$\begin{aligned}
&\mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right] \\
=&\Phi_{F_n,f}(\Xi_s^\infty) - \Phi_{F_n,f}(\Xi_0^\infty) - \int_0^s \mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty)du \\
&+ \mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{F_n,f}(\Xi_t^\infty)\middle|\mathcal{F}_s\right] + \mathbf{E}\left[\Phi_{F_n,f}(\Xi_0^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty)\middle|\mathcal{F}_s\right] \\
&+ \int_0^t \mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty) - \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)\middle|\mathcal{F}_s\right] du.
\end{aligned}$$

Since this is true for all $n \in \mathbb{N}^*$,

$$\begin{aligned}
&\mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)du \middle| \mathcal{F}_s\right] \\
=& \lim_{n\to+\infty} \Phi_{F_n,f}(\Xi_s^\infty) - \lim_{n\to+\infty} \Phi_{F_n,f}(\Xi_0^\infty) - \lim_{n\to+\infty} \int_0^s \mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty)du \\
&+ \lim_{n\to+\infty} \mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{F_n,f}(\Xi_t^\infty)\middle|\mathcal{F}_s\right] + \lim_{n\to+\infty} \mathbf{E}\left[\Phi_{F_n,f}(\Xi_0^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty)\middle|\mathcal{F}_s\right] \\
&+ \lim_{n\to+\infty} \int_0^t \mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty) - \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)\middle|\mathcal{F}_s\right] du,
\end{aligned}$$

under the condition that all these limits exist.

First, since $\Phi_{F_n,f}$ converges pointwise to $\Phi_{\delta_0,f}$,

$$\lim_{n\to+\infty} \Phi_{F_n,f}(\Xi_s^\infty) - \Phi_{F_n,f}(\Xi_0^\infty) = \Phi_{\delta_0,f}(\Xi_s^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty),$$

and by the dominated convergence theorem,

$$\lim_{n\to+\infty} \mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{F_n,f}(\Xi_t^\infty)\middle|\mathcal{F}_s\right] = \lim_{n\to+\infty} \mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_0^\infty) - \Phi_{F_n,f}(\Xi_0^\infty)\middle|\mathcal{F}_s\right] = 0.$$

Moreover, since for all $n \in \mathbb{N}^*$,

$$\int_0^s \left|\mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty)\right| du \leq 2s \times C_t \times \int_0^\infty \left(\mathcal{R}^d \vee 1\right) \mu(d\mathcal{R}),$$

again by the dominated convergence theorem, we obtain

$$\lim_{n\to+\infty} \int_0^s \mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty)du = \int_0^s \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)du.$$

Then, let $n \in \mathbb{N}^*$. Recalling that Ξ_u^∞ is also denoted $m(E_u)$,

$$\begin{aligned}
&\int_0^t \mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty) - \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)\middle|\mathcal{F}_s\right] du \\
=& \int_0^t \mathbf{E}\left[\int_0^\infty \int_{S^{\mathcal{R}}(E)} \left(F_n\left(\langle m(E_u \cup \mathcal{B}(x,\mathcal{R})), f\rangle\right) - \delta_0\left(\langle m(E_u \cup \mathcal{B}(x,\mathcal{R})), f\rangle\right)\right) dx\mu(d\mathcal{R})\middle|\mathcal{F}_s\right] du \\
&+ \int_0^t \mathbf{E}\left[\int_0^\infty \int_{S^{\mathcal{R}}(E)} \left(\delta_0\left(\langle m(E_u), f\rangle\right) - F_n\left(\langle m(E_u), f\rangle\right)\right) dx\mu(d\mathcal{R})\middle|\mathcal{F}_s\right] du.
\end{aligned}$$

Since for all $x \in \mathbb{R}^d$, $u \in [0,t]$ and $\mathcal{R} > 0$,

$$\lim_{n\to+\infty} F_n\left(\langle m(E_u \cup \mathcal{B}(x,\mathcal{R})), f\rangle\right) = \delta_0\left(\langle m(E_u \cup \mathcal{B}(x,\mathcal{R})), f\rangle\right)$$

and $$\lim_{n\to+\infty} F_n\left(\langle m(E_u), f\rangle\right) = \delta_0\left(\langle m(E_u), f\rangle\right),$$

using the dominated convergence theorem, we obtain that

$$\lim_{n\to+\infty}\int_0^t \mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F_n,f}(\Xi_u^\infty) - \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)\middle|\mathcal{F}_s\right] = 0,$$

and we can conclude that

$$\begin{aligned}&\mathbf{E}\left[\Phi_{\delta_0,f}(\Xi_t^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty(\Xi_0^\infty) - \int_0^t \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)du\middle|\mathcal{F}_s\right]\\ =&\Phi_{\delta_0,f}(\Xi_s^\infty) - \Phi_{\delta_0,f}(\Xi_0^\infty(\Xi_0^\infty) - \int_0^s \mathcal{G}_\mu^\infty \Phi_{\delta_0,f}(\Xi_u^\infty)du.\end{aligned}$$

□

3.4.3 Uniqueness of the solution to the martingale problem characterizing the ∞-parent SLFV

We now use the extended martingale problem in order to prove Proposition 3.1.13, i.e, that the ∞-parent ancestral process is the dual of the ∞-parent SLFV.

Proof. (Proposition 3.1.13) For all $t \geq 0$, let ω_t be a density of M_t^∞. Let $(E_t)_{t\geq 0}$ be such that

$$(\Xi_t^\infty)_{t\geq 0} = (m(E_t))_{t\geq 0}.$$

For all $s, t \geq 0$, we set :

$$F(s,t) = \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\widetilde{D}(M_s^\infty, \Xi_t^\infty)\right]\right].$$

Then,

$$F(s,t) = \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\Phi_{\delta_0,1-\omega_s}(\Xi_t^\infty)\right]\right]$$

and by Lemma 3.4.6,

$$F(s,t) = \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\Phi_{\delta_0,1-\omega_s}(\Xi_0)\right]\right] + \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\int_0^t \mathcal{G}_\mu^\infty \Phi_{\delta_0,1-\omega_s}(\Xi_u^\infty)du\right]\right].$$

By Fubini's theorem, we obtain

$$F(s,t) = F(s,0) + \int_0^t \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\mathcal{G}_\mu^\infty \Phi_{\delta_0,1-\omega_s}(\Xi_u^\infty)\right]\right]du.$$

Then,

$$\begin{aligned}F(s,t) &= \mathbf{E}_{m(E^0)}\left[\mathbb{E}_{M^0}\left[\widetilde{D}(M_s^\infty, \Xi_t^\infty)\right]\right]\\ &= \mathbf{E}_{m(E^0)}\left[\mathbb{E}_{M^0}\left[\Psi_{F^{\mathrm{Vol}(E_t)},f^{E_t}}(M_s^\infty)\right]\right],\end{aligned}$$

and by Lemma 3.4.1,

$$\begin{aligned}F(s,t) &= \mathbf{E}_{m(E^0)}\left[\mathbb{E}_{M^0}\left[\Psi_{F^{\mathrm{Vol}(E_t)},f^{E_t}}(M_0^\infty)\right]\right] + \mathbf{E}_{m(E^0)}\left[\mathbb{E}_{M^0}\left[\int_0^s \mathcal{L}_\mu^\infty \Psi_{\delta_{F^{\mathrm{Vol}(E_t)},f^{E_t}}}(M_u^\infty)du\right]\right]\\ &= F(0,t) + \mathbf{E}_{m(E^0)}\left[\mathbb{E}_{M^0}\left[\int_0^s \mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E_t)},f^{E_t}}(M_u^\infty)du\right]\right].\end{aligned}$$

Again by Fubini's theorem, we obtain

$$F(s,t) = F(0,t) + \int_0^s \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E_t)},f^{E_t}}(M_u^\infty)\right]\right] du.$$

Combining both expressions for $F(s,t)$, by Lemma 4.4.10 in [EK86], we obtain :

$$\begin{aligned}
&F(t,0) - F(0,t)\\
&= \int_0^t \left(\mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E_{t-u})},f^{E_{t-u}}}(M_u^\infty)\right]\right] - \mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\mathcal{G}_\mu^\infty \Phi_{\delta_0,1-\omega_u}(\Xi_{t-u}^\infty)\right]\right]\right) du.
\end{aligned}$$

Let $u \in [0,t]$. We have

$$\begin{aligned}
&\mathcal{G}_\mu^\infty \Phi_{\delta_0,1-\omega_u}(\Xi_{t-u}^\infty)\\
&= \int_0^\infty \int_{S^{\mathcal{R}}(E_{t-u})} \left(\delta_0\left(\int_{E_{t-u}\cup\mathcal{B}(x,\mathcal{R})} (1-\omega_u(z))\,dz\right) - \delta_0\left(\int_{E_{t-u}} (1-\omega_u(z))\,dz\right)\right) dx\mu(d\mathcal{R})
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{L}_\mu^\infty \Psi_{F^{\mathrm{Vol}(E_{t-u})},f^{E_{t-u}}}(M_u^\infty)\\
&= \int_0^\infty \int_{S^{\mathcal{R}}(E_{t-u})} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_u(z))\,dz\right)\right)\\
&\quad \times \left[\delta_0\left(\mathrm{Vol}(E_{t-u}) - \langle \Theta_{x,\mathcal{R}}^-(\omega_u), \mathbb{1}_{E_{t-u}}\rangle\right) - \delta_0\left(\mathrm{Vol}(E_{t-u}) - \langle \omega_u, \mathbb{1}_{E_{t-u}}\rangle\right)\right] dx\\
&= \int_0^\infty \int_{S^{\mathcal{R}}(E_{t-u})} \left(1 - \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_u(z))\,dz\right)\right)\\
&\quad \times \left[\delta_0\left(\mathrm{Vol}(E_{t-u}) - \langle \Theta_{x,\mathcal{R}}^-(\omega_u), \mathbb{1}_{E_{t-u}}\rangle\right) - \delta_0\left(\int_{E_{t-u}} (1-\omega_u(z))\,dz\right)\right] dx.
\end{aligned}$$

For all $\mathcal{R} > 0$ and $x \in S^{\mathcal{R}}(E_{t-u})$,

$$\begin{aligned}
\delta_0\left(\mathrm{Vol}(E_{t-u}) - \langle \Theta_{x,\mathcal{R}}^-(\omega_u), \mathbb{1}_{E_{t-u}}\rangle\right) &= \delta_0\left(\mathrm{Vol}(E_{t-u}) - \int_{E_{t-u}\setminus\mathcal{B}(x,\mathcal{R})} \omega_u(z)dz\right)\\
&= \delta_0\left(\mathrm{Vol}(E_{t-u}\cap\mathcal{B}(x,\mathcal{R})) + \int_{E_{t-u}\setminus\mathcal{B}(x,\mathcal{R})} (1-\omega_u(z))\,dz\right).
\end{aligned}$$

Since $x \in S^{\mathcal{R}}(E_{t-u})$, $\mathrm{Vol}(E_{t-u}\cap\mathcal{B}(x,\mathcal{R})) \neq 0$, and hence

$$\delta_0\left(\mathrm{Vol}(E_{t-u}) - \langle \Theta_{x,\mathcal{R}}^-(\omega_u), \mathbb{1}_{E_{t-u}}\rangle\right) = 0.$$

Moreover, notice that

$$\delta_0\left(\int_{E_{t-u}\cup\mathcal{B}(x,\mathcal{R})} (1-\omega_u(z))\,dz\right) = \delta_0\left(\int_{E_{t-u}} (1-\omega_u(z))\,dz\right) \times \delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega_u(z))\,dz\right).$$

Therefore,

$$\begin{aligned}
&\mathcal{L}_{\mu}^{\infty}\Psi_{F^{\mathrm{Vol}(E_{t-u})},f^{E_{t-u}}}(M_u^{\infty})\\
&=\int_0^{\infty}\int_{S^{\mathcal{R}}(E_{t-u})}\delta_0\left(\int_{\mathcal{B}(x,\mathcal{R})}(1-\omega_u(z))\,dz\right)\times\delta_0\left(\int_{E_{t-u}}(1-\omega_u(z))\,dz\right)dx\mu(d\mathcal{R})\\
&\quad-\int_0^{\infty}\int_{S^{\mathcal{R}}(E_{t-u})}\delta_0\left(\int_{E_{t-u}}(1-\omega_u(z))\,dz\right)dx\mu(d\mathcal{R})\\
&=\int_0^{\infty}\int_{\mathbb{R}^d}\mathbb{1}_{x\in S^{\mathcal{R}}(E_{t-u})}\times\left[\delta_0\left(\int_{E_{t-u}\cup\mathcal{B}(x,\mathcal{R})}(1-\omega_u(z))\,dz\right)-\delta_0\left(\int_{E_{t-u}}(1-\omega_u(z))\,dz\right)\right]dx\mu(d\mathcal{R}),
\end{aligned}$$

which is equal to $\mathcal{G}_{\mu}^{\infty}\Phi_{\delta_0,1-\omega_u}(\Xi_{t-u}^{\infty})$. Thus

$$F(t,0)=F(0,t)$$

i.e
$$\mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\widetilde{D}(M_t^{\infty},\Xi_0^{\infty})\right]\right]=\mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\widetilde{D}(M_0^{\infty},\Xi_t^{\infty})\right]\right]$$

$$\iff\quad\mathbb{E}_{M^0}\left[\mathbf{E}_{m(E^0)}\left[\delta_0\left(\int_{E_0}(1-\omega_t(x))\,dx\right)\right]\right]=\mathbb{E}_{M^0}\left[\mathbf{E}_{m(E_t)}\left[\delta_0\left(\int_{E_0}(1-\omega_0(x))\,dx\right)\right]\right].$$

Therefore

$$\mathbb{E}_{M^0}\left[\delta_0\left(\int_{E_0}(1-\omega_t(x))\,dx\right)\right]=\mathbf{E}_{m(E^0)}\left[\delta_0\left(\int_{E_t}(1-\omega_0(x))\,dx\right)\right]$$

and we can conclude. □

Finally, we can prove the second part of Theorem 3.1.10, i.e, the uniqueness of the solution to the martingale problem satisfied by the ∞-parent SLFV when μ satisfies Condition (3.1.5). The first part of this theorem was proved in Section 3.3 (Proposition 3.2.10).

Proof. (Theorem 3.1.10)

Let $(M_t^1)_{t\geq 0}$ and $(M_t^2)_{t\geq 0}$ be two solutions to the martingale problem $(\mathcal{L}_{\mu}^{\infty},\delta_{M^0})$ with values in $\mathcal{M}_{\lambda}$. Then, there exists densities $(\omega_t^1)_{t\geq 0}$ and $(\omega_t^2)_{t\geq 0}$ of $(M_t^1)_{t\geq 0}$ and $(M_t^2)_{t\geq 0}$ such that

$$\forall t\geq 0,\forall x\in\mathbb{R}^d,\omega_t^1(x)\in\{0,1\}\text{ and }\omega_t^2(x)\in\{0,1\}.$$

Then, let $t\geq 0$, let $E\in\mathcal{E}^{cf}$ and let $(\Xi_t^{\infty})_{t\geq 0}$ be the ∞-parent ancestral process started from $m(E)$. We have

$$\begin{aligned}
\mathbb{P}_{M^0}\left(\delta_0\left(\int_E\left(1-\omega_t^1(x)\right)dx\right)=1\right)&=\mathbb{E}_{M^0}\left[\delta_0\left(\int_E\left(1-\omega_t^1(x)\right)dx\right)\right]\\
&=\mathbb{E}_{M^0}\left[\widetilde{D}(M_{\infty}^1,\Xi_0^{\infty})\right]\\
&=\mathbf{E}_{m(E)}\left[\widetilde{D}(M^0,\Xi_t^{\infty})\right]\text{ by Proposition 3.1.13}\\
&=\mathbb{E}_{M^0}\left[\widetilde{D}(M_{\infty}^2,\Xi_0^{\infty})\right]\text{ by the same proposition}\\
&=\mathbb{E}_{M^0}\left[\delta_0\left(\int_E\left(1-\omega_t^2(x)\right)dx\right)\right]\\
&=\mathbb{P}_{M^0}\left(\delta_0\left(\int_E\left(1-\omega_t^2(x)\right)dx\right)=1\right),
\end{aligned}$$

using Proposition 3.1.13 to pass from line 2 to line 3, and from line 3 to line 4. Since ω_t^1 and ω_t^2 are $\{0,1\}$-valued, this implies that the supports of M_t^1 and M_t^2 are equal in distribution for all $t\geq 0$, which allows us to conclude that $(M_t^1)_{t\geq 0}$ and $(M_t^2)_{t\geq 0}$ have the same distribution. □

3.5 Technical lemmas

3.5.1 Properties of the operators $\mathcal{L}_\mu^k$ and $\mathcal{L}_\mu^\infty$

The goal of this section is to show that the operators $\mathcal{L}_\mu^k$ and $\mathcal{L}_\mu^\infty$ introduced in Section 3.1 are well-defined, as well as to prove some properties they satisfy.

In all that follows, let $F \in C^1(\mathbb{R})$, $f \in C_c(\mathbb{R}^d)$, and $M \in \mathcal{M}_\lambda$. Let $\omega : \mathbb{R}^d \to \{0,1\}$ be a measurable function, let μ be a σ-finite measure on $\mathbb{R}_+^*$ satisfying Condition (3.1.4), and let $k \geq 2$. Since f is of compact support, there exist constants $C_1, C_2 > 0$ such that for all $\mathcal{R} > 0$,

$$\mathrm{Vol}(Supp^{\mathcal{R}}(f)) \leq C_2 \times \left(\mathcal{R}^d \vee 1\right), \tag{3.5.1}$$

and for all $\tilde{\omega} : \mathbb{R}^d \to \{0,1\}$ measurable,

$$\left|\langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})} \times \tilde{\omega}, f\rangle\right| \leq C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right). \tag{3.5.2}$$

Lemma 3.5.1. *For all $x \in \mathbb{R}^d$ and for all $\mathcal{R} > 0$,*

$$\left|\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right| \leq ||f||_\infty \times \mathrm{Vol}(Supp(f))$$

and $$\left|\langle \Theta^-_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right| \leq ||f||_\infty \times \mathrm{Vol}(Supp(f)).$$

Proof. Let $x \in \mathbb{R}^d$ and $\mathcal{R} > 0$.

$$\begin{aligned}
\left|\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right| &\leq \left|\langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})^c} \times \omega, f\rangle + \langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})}, f\rangle - \langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})^c} \times \omega, f\rangle - \langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})} \times \omega, f\rangle\right| \\
&\leq \left|\langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})} \times (1-\omega), f\rangle\right| \\
&\leq \left|\int_{\mathcal{B}(x,\mathcal{R})} (1-\omega(y)) \times f(y) dy\right| \\
&\leq \int_{\mathcal{B}(x,\mathcal{R})} |f(y)| dy \\
&\leq ||f||_\infty \times \mathrm{Vol}(Supp(f)).
\end{aligned}$$

We can similarly show the corresponding result for $\left|\langle \Theta^-_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right|$. □

Lemma 3.5.2. *For all $\mathcal{R} > 0$, for all $x \in \mathbb{R}^d \backslash Supp^{\mathcal{R}}(f)$,*

$$\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle = \langle \Theta^-_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle = 0$$

Proof. Let $\mathcal{R} > 0$, and let $x \in \mathbb{R}^d \backslash Supp^{\mathcal{R}}(f)$,

$$\begin{aligned}
\left|\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right| &= \left|\langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})} \times (1-\omega), f\rangle\right| \\
&\leq \int_{\mathcal{B}(x,\mathcal{R})} |f(y)| dy \\
&= 0
\end{aligned}$$

since $x \in \mathbb{R}^d \backslash Supp^{\mathcal{R}}(f)$. Similarly,

$$\begin{aligned}
\left|\langle \Theta^-_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right| &= \left|\langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})} \times \omega, f\rangle\right| \\
&\leq \int_{\mathcal{B}(x,\mathcal{R})} |f(y)| dy \\
&= 0
\end{aligned}$$

for the same reason, and we can conclude. □

Lemma 3.5.3. *For all $x \in \mathbb{R}^d$ and for all $\mathcal{R} > 0$,*

$$\begin{aligned} \left|F\left(\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle\right) - F\left(\langle \omega, f\rangle\right)\right| &\leq C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) \\ \textit{and} \quad \left|F\left(\langle \Theta^-_{x,\mathcal{R}}(\omega), f\rangle\right) - F\left(\langle \omega, f\rangle\right)\right| &\leq C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) \end{aligned}$$

where

$$C(F,f) = \sup_{z \in [-||f||_\infty \mathrm{Vol}(Supp(f)), ||f||_\infty \mathrm{Vol}(Supp(f))]} \left|F'(z)\right|.$$

Proof. Let $x \in \mathbb{R}^d$ and $\mathcal{R} > 0$. First, we notice that as in the proof of Lemma 3.5.1, we only need to show the result for $\Theta^+_{x,\mathcal{R}}(\omega)$.

By Taylor-Lagrange inequality and by Lemma 3.5.1,

$$\begin{aligned} \left|F\left(\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle\right) - F\left(\langle \omega, f\rangle\right)\right| &\leq \left|\langle \Theta^+_{x,\mathcal{R}}(\omega), f\rangle - \langle \omega, f\rangle\right| \times C(F,f) \\ &\leq \left|\langle \mathbb{1}_{\mathcal{B}(x,\mathcal{R})} \times (1-\omega), f\rangle\right| \times C(F,f) \\ &\leq C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) \end{aligned}$$

by Eq. (3.5.2). □

We can now show that the operator $\mathcal{L}^k_\mu$ is well-defined.

Lemma 3.5.4. *The operator $\mathcal{L}^k_\mu$ is well-defined. Moreover, the function $\mathcal{L}^k_\mu \Psi_{F,f} : \mathcal{M}_\lambda \to \mathbb{R}$ is bounded.*

Proof. Let $M \in \mathcal{M}_\lambda$. Then

$$\begin{aligned} &\left|\mathcal{L}^k_\mu \Psi_{F,f}(M)\right| \\ &\leq \left|\int_{\mathbb{R}^d} \int_0^\infty \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V^k_\mathcal{R}} \left(\prod_{j=1}^k \omega_M(y_j)\right) \left(F\left(\langle \Theta^+_{x,\mathcal{R}}(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dy_1 ... dy_k \mu(d\mathcal{R}) dx\right| \\ &\quad + \left|\int_{\mathbb{R}^d} \int_0^\infty \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V^k_\mathcal{R}} \left(1 - \prod_{j=1}^k \omega_M(y_j)\right) \left(F\left(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dy_1 ... dy_k \mu(d\mathcal{R}) dx\right| \\ &\leq \left|\int_{\mathbb{R}^d} \int_0^\infty \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V^k_\mathcal{R}} \times \left(F\left(\langle \Theta^+_{x,\mathcal{R}}(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dy_1 ... dy_k \mu(d\mathcal{R}) dx\right| \\ &\quad + \left|\int_{\mathbb{R}^d} \int_0^\infty \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V^k_\mathcal{R}} \times \left(F\left(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dy_1 ... dy_k \mu(d\mathcal{R}) dx\right| \\ &\leq \left|\int_0^\infty \int_{Supp^\mathcal{R}(f)} \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V^k_\mathcal{R}} \times \left(F\left(\langle \Theta^+_{x,\mathcal{R}}(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dy_1 ... dy_k dx \mu(d\mathcal{R})\right| \\ &\quad + \left|\int_0^\infty \int_{Supp^\mathcal{R}(f)} \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V^k_\mathcal{R}} \times \left(F\left(\langle \Theta^-_{x,\mathcal{R}}(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dy_1 ... dy_k dx \mu(d\mathcal{R})\right|. \end{aligned}$$

Using Lemma 3.5.3,

$$\begin{aligned} &\left|\mathcal{L}^k_\mu \Psi_{F,f}(M)\right| \\ &\leq \int_0^\infty \int_{Supp^\mathcal{R}(f)} \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{2}{V^k_\mathcal{R}} \times C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) dy_1 ... dy_k dx \mu(d\mathcal{R}). \\ &\leq \int_0^\infty 2\mathrm{Vol}(Supp^\mathcal{R}(f)) \times C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) \mu(d\mathcal{R}), \end{aligned}$$

and by Eq. (3.5.1),

$$\begin{aligned}
&\left|\mathcal{L}_\mu^k \Psi_{F,f}(M)\right| \\
\leq &\int_0^\infty 2C_1C_2 \times ||f||_\infty \times C(F,f) \times \left(\mathcal{R}^d \wedge 1\right) \times \left(\mathcal{R}^d \vee 1\right) \mu(d\mathcal{R}) \\
\leq &2C_1C_2||f||_\infty \times C(F,f) \times \int_0^\infty \mathcal{R}^d \mu(d\mathcal{R}) \\
< &+\infty
\end{aligned}$$

since μ satisfies Condition (3.1.4).

The second part of the lemma is a direct consequence of the fact that

$$2C_1C_2||f||_\infty \times C(F,f) \times \int_0^\infty \mathcal{R}^d \mu(d\mathcal{R})$$

does not depend on the choice of M. □

A consequence of this lemma and of Lemma 3.5.2 is that for all $M \in \mathcal{M}_\lambda$, $\mathcal{L}_\mu^k \Psi_{F,f}(M)$ can be rewritten as :

$$\begin{aligned}
\mathcal{L}_\mu^k \Psi_{F,f}(M) = \int_0^\infty \int_{Supp^{\mathcal{R}}(f)} \int_{\mathcal{B}(x,\mathcal{R})^k} \frac{1}{V_{\mathcal{R}}^k} \times \Big[&\prod_{j=1}^k \omega_M(y_j) \times F(\langle \Theta_{x,\mathcal{R}}^+(\omega_M), f\rangle) \\
&+ (1 - \prod_{j=1}^k \omega_M(y_j)) \times F(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f\rangle) \\
&- F(\langle \omega_M, f\rangle)\Big] dy_1 ... dy_k dx \mu(d\mathcal{R}).
\end{aligned}$$

We now prove that the operator $\mathcal{L}_\mu^\infty$ is well-defined.

Lemma 3.5.5. *The operator $\mathcal{L}_\mu^\infty$ is well-defined. Moreover, the function $\mathcal{L}_\mu^\infty \Psi_{F,f} : \mathcal{M}_\lambda \to \mathbb{R}$ is bounded.*

Proof. Let $M \in \mathcal{M}_\lambda$. Then,

$$\begin{aligned}
&\left|\mathcal{L}_\mu^\infty \Psi_{F,f}(M)\right| \\
\leq &\int_0^\infty \int_{Supp^{\mathcal{R}}(f)} \left| \left(1 - \delta_0 \left(\int_{\mathcal{B}(x,\mathcal{R})} 1 - \omega_M(z) dz\right)\right) \times \left[F\left(\langle \Theta_{x,\mathcal{R}}^-(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right] \right| dx \mu(d\mathcal{R}) \\
\leq &\int_0^\infty \int_{Supp^{\mathcal{R}}(f)} C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) dx \mu(d\mathcal{R}) \\
\leq &\int_0^\infty Vol(Supp^{\mathcal{R}}(f)) C_1 \times ||f||_\infty \times \left(\mathcal{R}^d \wedge 1\right) \times C(F,f) dx \mu(d\mathcal{R}) \\
\leq &C_1C_2C(F,f) \times ||f||_\infty \times \int_0^\infty \left(\mathcal{R}^d \wedge 1\right) \times \left(\mathcal{R}^d \vee 1\right) \mu(d\mathcal{R}) \\
\leq &C_1C_2C(F,f) \times ||f||_\infty \times \int_0^\infty \mathcal{R}^d \mu(d\mathcal{R}) \\
< &+\infty
\end{aligned}$$

since μ satisfies Condition (3.1.4). Here we used Lemma 3.5.3 to pass from the second to the third line, and Lemma 3.5.1 to pass from the fourth to the fifth line.

As before, the second part of the lemma is the consequence of the fact that

$$C_1C_2C(F,f)\times||f||_\infty\times\int_0^\infty\mathcal{R}^d\mu(d\mathcal{R})$$

does not depend on the choice of M. □

3.5.2 Properties of the operator $\mathcal{G}_\mu^\infty$

In all the following, let μ be a σ-finite measure on $\mathbb{R}_+^*$ satisfying Condition (3.1.5), let $F\in C_b^1(\mathbb{R})$ and let $f\in\mathcal{B}(\mathbb{R}^d)$.

Lemma 3.5.6. *The operator $\mathcal{G}_\mu^\infty$ is well-defined, and the function $\mathcal{G}_\mu^\infty\Phi_{F,f}$ is bounded. In particular, $\Phi_{F,f}$ belongs to the domain of the operator $\mathcal{G}_\mu^\infty$.*

Proof. Let $m(E)\in\mathcal{M}^{cf}$. Then,

$$\begin{aligned}
\left|\mathcal{G}_\mu^\infty\Phi_{F,f}(m(E))\right|\leq&\int_0^\infty\int_{S^{\mathcal{R}}(E)\cap Supp^{\mathcal{R}}(f)}\left|F\left(\langle m(E\cup\mathcal{B}(x,\mathcal{R})),f\rangle\right)-F\left(\langle m(E),f\rangle\right)\right|dx\mu(d\mathcal{R})\\
\leq&\int_0^\infty\int_{S^{\mathcal{R}}(E)\cap Supp^{\mathcal{R}}(f)}2||F||_\infty dx\mu(d\mathcal{R})\\
\leq&2||F||_\infty\int_0^\infty\mathrm{Vol}(S^{\mathcal{R}}(E)\cap Supp^{\mathcal{R}}(f))\mu(d\mathcal{R})\\
\leq&2||F||_\infty\int_0^\infty\mathrm{Vol}(Supp^{\mathcal{R}}(f))\mu(d\mathcal{R})\\
\leq&2||F||_\infty\int_0^\infty C_2\times\left(\mathcal{R}^d\vee 1\right)\mu(d\mathcal{R})\\
<&+\infty,
\end{aligned}$$

since μ satisfies Condition (3.1.4). □

Lemma 3.5.7. *Let $\Xi\in\mathcal{M}^{cf}$, and let $(\Xi_t)_{t\geq0}$ be the ∞-parent ancestral process associated to μ with initial condition Ξ. Then, for all $t\geq0$,*

$$\mathbf{E}\left[\int_0^t\mathcal{G}_\mu^\infty\Phi_{F,f}(\Xi_s)ds\right]=\int_0^t\mathbf{E}\left[\mathcal{G}_\mu^\infty\Phi_{F,f}(\Xi_s)\right]ds.$$

Proof. Let $t\geq0$, and let $\tilde{\mathcal{R}}>0$ such that μ satisfies Condition (3.1.5). Then, since $u\to\mathrm{Vol}(\Xi_u)$ is increasing,

$$\begin{aligned}
\mathbf{E}\left[\int_0^t\left|\mathcal{G}_\mu^\infty\Phi_{F,f}(\Xi_s)\right|ds\right]\leq&2||F||_\infty\times\mathbf{E}\left[\int_0^t\int_0^\infty\mathrm{Vol}(S^{\mathcal{R}}(\Xi_s))\mu(d\mathcal{R})ds\right]\\
\leq&2||F||_\infty\times\mathbf{E}\left[\int_0^t\int_0^\infty\mathrm{Vol}(S^{\mathcal{R}}(\Xi_t))\mu(d\mathcal{R})ds\right]\\
\leq&2||F||_\infty\times t\times\mathbf{E}\left[\int_0^\infty C_t\times\mathrm{Vol}(\mathcal{B}(0,\mathcal{R}))\mu(d\mathcal{R})\right],
\end{aligned}$$

with $(C_t)_{t\geq0}$ being the $\tilde{\mathcal{R}}$-covering process associated to $(\Xi_t)_{t\geq0}$. Therefore,

$$\begin{aligned}
\mathbf{E}\left[\int_0^t\left|\mathcal{G}_\mu^\infty\Phi_{F,f}(\Xi_s)\right|ds\right]\leq&2||F||_\infty\times t\times\int_0^\infty V_1\times\left(\tilde{\mathcal{R}}+\mathcal{R}\right)^d\mu(d\mathcal{R})\times\mathbf{E}\left[C_t\right]\\
\leq&2||F||_\infty\times t\times\int_0^\infty V_1\times\left(\tilde{\mathcal{R}}+\mathcal{R}\right)^d\mu(d\mathcal{R})\times\mathbf{E}\left[Y_t\right],
\end{aligned}$$

where Y_t is the branching process associated to $(C_t)_{t\geq 0}$ introduced in Section 3.3. Hence

$$\mathbf{E}\left[\int_0^t \left|\mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_s)\right| ds\right] < +\infty.$$

We conclude by applying Fubini's theorem. □

Lemma 3.5.8. *Let $\Xi \in \mathcal{M}^{cf}$, and let $(\Xi_t)_{t\geq 0}$ be the ∞-parent ancestral process associated to μ with initial condition Ξ. Then, for all $t \geq 0$,*

$$\mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_t)\right] = \frac{d}{du}\mathbf{E}\left[\Phi_{F,f}(\Xi_u)\right]\Big|_{u=t}.$$

Proof. Let $t \geq 0$. By Lemma 3.5.6, $\Phi_{F,f}$ is in the domain of $\mathcal{G}_\mu^\infty$. Moreover, by considering the number of jumps of $(\Xi_t)_{t\geq 0}$ over the time interval $[0,\epsilon]$, $\epsilon \geq 0$ and controlling their size, it is straightforward to show that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathbb{E}\left[\Phi_{F,f}(\Xi_t)\right] - \Phi_{F,f}(\Xi_0)\right) = \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_t),$$

or in other words, that

$$\frac{d}{du}\mathbf{E}\left[\Phi_{F,f}(\Xi_t)\right]\Big|_{t=0} = \mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_t).$$

Therefore, for all $s \in [0,t]$,

$$\mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_s)\right] = \mathbf{E}\left[\frac{d}{du}\mathbf{E}\left[\Phi_{F,f}(\Xi_t)|\Xi_s\right]\Big|_{t=s}\right].$$

Since F' is bounded, by the dominated convergence theorem,

$$\begin{aligned}\mathbf{E}\left[\mathcal{G}_\mu^\infty \Phi_{F,f}(\Xi_s)\right] &= \frac{d}{du}\mathbf{E}\left[\mathbf{E}\left[\Phi_{F,f}(\Xi_t)|\Xi_s\right]\right]\Big|_{t=s} \\ &= \frac{d}{du}\mathbf{E}\left[\Phi_{F,f}(\Xi_t)\right]\Big|_{t=s}\end{aligned}$$

and we can conclude. □

3.5.3 Properties of the densities of coupled k-parent SLFVs

The goal of this section is to prove technical lemmas about the density of coupled k-parent SLFVs, which will be used in Section 3.3 in order to construct the ∞-parent SLFV.

In all that follows, let μ be a σ-finite measure on $(0,+\infty)$ satisfying Condition (3.1.4), and let Π^c be a Poisson point process on $\mathbb{R}\times\mathbb{R}^d\times(0,+\infty)\times U$ with intensity

$$dt \otimes dx \otimes \mu(d\mathcal{R}) \otimes \tilde{u}\left(d(p_n)_{n\geq 1}\right).$$

Lemma 3.5.9. *For all $k \geq 2$, for all $0 \leq s \leq t$ and for all $x \in \mathbb{R}^d$,*

$$A\left(\Xi_{k,t}^{\Pi^c,t,\delta_x}\right) = \bigcup_{x'\in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)} A\left(\Xi_{k,s}^{\Pi^c,s,\delta_{x'}}\right).$$

Proof. Let $k \geq 2$, let $0 \leq s \leq t$ and let $x \in \mathbb{R}^d$. Let $y \in A\left(\Xi_{k,t}^{\Pi^c,t,\delta_x}\right)$. Then, we can construct a chain of reproduction events linking the point x at time t to the point y at time 0. We can split it into two chains :

- one linking the point x at time t to a point $y' \in \mathbb{R}^d$ at time s,
- one linking the point y' at time s to the point y at time 0.

Therefore, $y' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)$ and $y \in A\left(\Xi_{k,s}^{\Pi^c,s,\delta_{y'}}\right)$, which means that

$$y \in \bigcup_{x' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)} A\left(\Xi_{k,s}^{\Pi^c,s,\delta_{x'}}\right).$$

Conversely, let y belonging to this set. It means that there exists $x' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)$ such that $y \in A\left(\Xi_{k,s}^{\Pi^c,s,\delta_{x'}}\right)$. Therefore, we can construct two chains of reproduction events, linking the point x at time t to the point x' at time s, and the point x' at time s to the point y at time 0. Hence $y \in A\left(\Xi_{k,t}^{\Pi^c,t,\delta_x}\right)$, and we can conclude. □

Lemma 3.5.10. *For all $k \geq 2$, for all $0 \leq s \leq t$ and for all $x \in \mathbb{R}^d$,*

$$\omega_{k,t}^{\Pi,\omega}(x) = \prod_{x' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)} \omega_{k,s}^{\Pi^c,\omega}(x').$$

Proof. Let $k \geq 2$, let $0 \leq s \leq t$ and let $x \in \mathbb{R}^d$. By definition,

$$\omega_{k,t}^{\Pi^c,\omega}(x) = \prod_{y \in A\left(\Xi_{k,t}^{\Pi^c,t,\delta_x}\right)} \omega(y) \tag{3.5.3}$$

and

$$\prod_{x' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)} \omega_{k,s}^{\Pi^c,\omega}(x') = \prod_{x' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)} \prod_{y \in A\left(\Xi_{k,s}^{\Pi^c,s,\delta_{x'}}\right)} \omega(y). \tag{3.5.4}$$

Since by Lemma 3.5.9

$$A\left(\Xi_{k,t}^{\Pi^c,t,\delta_x}\right) = \bigcup_{x' \in A\left(\Xi_{k,t-s}^{\Pi^c,t,\delta_x}\right)} A\left(\Xi_{k,s}^{\Pi^c,s,\delta_{x'}}\right),$$

the same terms appear in both products. However, some terms may appear more than once in Eq. (3.5.4), while they can appear only once in Eq. (3.5.3). But ω is $\{0,1\}$-valued, so for all $y \in \mathbb{R}^d$ and $j \in \mathbb{N}^*$, $\omega^j(y) = \omega(y)$, and we can conclude. □

Lemma 3.5.11. *For all $\tilde{k} \geq 2$, for all $0 \leq s \leq t$ and for all $x \in \mathbb{R}^d$,*

$$\lim_{k \to +\infty} \prod_{x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)} \omega_{k,s}^{\Pi^c,\omega}(x') \leq \prod_{x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)} \lim_{k \to +\infty} \omega_{k,s}^{\Pi^c,\omega}(x').$$

Proof. Let $\tilde{k} \geq 2$, let $0 \leq s \leq t$ and let $x \in \mathbb{R}^d$. Since both quantities are $\{0,1\}$-valued, we only need to show that if

$$\prod_{x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)} \lim_{k\to+\infty} \omega_{k,s}^{\Pi^c,\omega}(x') = 0$$

then

$$\lim_{k\to+\infty} \prod_{x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)} \omega_{k,s}^{\Pi^c,\omega}(x') = 0.$$

Assume that the first equality is true. Then, there exists $x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)$ such that

$$\lim_{k\to+\infty} \omega_{k,s}^{\Pi^c,\omega}(x') = 0.$$

But since $\left(\omega_{k,s}^{\Pi^c,\omega}(x')\right)_{k\geq 2}$ is decreasing and $\{0,1\}$-valued, there exists $k' \geq 2$ such that for all $k \geq k'$, $\omega_{k,s}^{\Pi^c,\omega}(x') = 0$. Therefore, for all $k \geq k'$,

$$\prod_{x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)} \omega_{k,s}^{\Pi^c,\omega}(x') = 0,$$

which means that

$$\lim_{k\to+\infty} \prod_{x' \in A\left(\Xi_{\tilde{k},t-s}^{\Pi^c,t,\delta_x}\right)} \omega_{k,s}^{\Pi^c,\omega}(x') = 0.$$

□

Chapter 4

Growth properties of the ∞-parent spatial Lambda-Fleming Viot process

This chapter is based on a a joint work with Amandine Véber. The corresponding article is available on Arxiv [LV22], and was submitted to Electronic Journal of Probability.

Abstract

The ∞-parent spatial Lambda-Fleming Viot process, or ∞-parent SLFV, is a model for spatially expanding populations in which empty areas are filled with "ghost" individuals. The interest of this process lies in the fact that it is akin to a continuous-space version of the classical Eden growth model, while being associated to a dual process encoding genealogies and allowing one to study the evolution of the genetic diversity in such a population.

In this article, we focus on the growth properties of the ∞-parent SLFV, and compare them to those of other stochastic growth processes, such as the Eden model. In order to do so, we first define what can be interpreted as the speed of growth of the area covered with the subpopulation of real individuals. Using the associated dual process and a comparison with a first-passage percolation problem, we show that the growth of the region covered with real individuals in the ∞-parent SLFV is linear in time. We use numerical simulations to approximate the speed of growth, and conjecture that due to the growth dynamics at the front, this speed is higher than the expected speed from simple first-moment calculations.

We then study a toy model of two interacting growing piles of cubes in order to understand how the growth dynamics at the front edge can increase the global speed of growth of the "occupied" region. We obtain an explicit formula for this speed of growth in our toy model, using the invariant distribution of a discretized version of the model. This study is of interest on its own right, and its implications are not restricted to the case of the ∞-parent SLFV.

4.1 Introduction

The ∞-parent spatial Λ-Fleming Viot process, or ∞-parent SLFV, was introduced in [Lou21] as a model for expanding populations in $\mathbb{R}^2$. Its main feature is the use of "ghost" individuals (thereafter referred to as "type 0" individuals) to fill empty areas, adapting ideas from [DF16; HN08]. In [DF16; HN08; Lou21], ghost individuals can reproduce as well, modeling stochastic fluctuations in population sizes, but with a selective disadvantage against real individuals (thereafter referred to as "type 1" individuals), ensuring population expansions can indeed occur. The ∞-parent SLFV corresponds to the limit of the process introduced in [Lou21] when the selective advantage of type 1 individuals over type 0 individuals becomes infinitely strong. Therefore, in the limiting regime, we no longer observe local extinctions due to the stochasticity in reproduction at the front, where densities in type 1 individuals are lower.

One of the main interests of the ∞-parent SLFV lies in the fact that the growth of the area occupied by type 1 individuals corresponds to a continuous space version of the Eden growth model [Ede61], while the process in itself can be interpreted as a population genetics model based on some form of very strong natural selection. Although there exist other variants of the Eden growth model which are continuous in space [WLB95], the ∞-parent SLFV is equipped with tools allowing one to investigate genetic diversity patterns, assuming that real individuals are further subdivided into different types (for instance using the concept of "tracers", see [DF16; HN08]). Therefore, it is potentially suited to study how the characteristic genetic diversity patterns observed in real expanding populations [Gra+13; Hal+07; HN10] arise. In this paper, we will only focus on the region covered by real individuals, without specifying different subtypes of real (or type 1) individuals. However, understanding the family structure in this model is a first step towards the understanding of how genetic diversity evolves in a multitype population.

Compared with other population genetics models with selection and a spatial structure, in which selection is often taken to be weak, the parameter regime of interest in the ∞-parent SLFV with ghost and real individuals is different: here we are in a strong selection limit, motivated by the interpretation of ghost individuals as modeling empty areas. See e.g [EFP17; Eth+17; EFS17; EVY20; FP17] for examples of population genetics models with selection belonging to the same family of spatial Λ-Fleming Viot processes as the ∞-parent SLFV, which are studied under what can be considered as a weak or moderately weak selection limit.

As for these other Λ-Fleming Viot processes (see also [BEV10; EV12]), the ∞-parent SLFV differs from other population dynamics models by having reproduction driven by an exogeneous Poisson point process of *reproduction events* rather than by individual reproduction. These reproduction events affect all or a fraction of individuals in a specific area. In the original ∞-parent SLFV process, as well as in other spatial Λ-Fleming Viot processes, the affected area is a ball, whose radius can be fixed or random. However, biological experiments suggest that the shape of the area impacted by a reproduction event influences the observed genetic diversity patterns [Hal+07]. Moreover, this excludes the case of a preferential expansion direction, for example towards a resource by a phenomenon similar to chemotaxis. Therefore, in this article, we consider a variant of the ∞-parent SLFV in which the area affected by a reproduction event can be shaped like an ellipse rather than a ball.

Informally, the ∞-parent SLFV is constructed as follows. At any time t, the population is represented by a "density" $\omega_t : \mathbb{R}^2 \to \{0, 1\}$ such that for all $x \in \mathbb{R}^2$, $\omega_t(x) = 1$ if site x is empty at time t (i.e, filled with only ghost individuals), and $\omega_t(x) = 0$ if site x is occupied at time t (i.e, filled with only real individuals). Hence, ω_t is the indicator function of the empty area, and $1 - \omega_t$ is the indicator function of the area occupied by the population of real individuals. This convention was motivated by the duality formula proved in [Lou21], which we recall in Section 4.2. Then, each reproduction event affects a specific area, whose center is sampled randomly, and whose shape can be prescribed

in advance or be sampled according to some predetermined probability law. For the moment, to ease the exposition, let us consider that all reproduction events have the same shape: a ball with fixed radius $\mathcal{R} > 0$. In Section 4.2, we will consider the more general case of ellipses with random bounded parameters.

Let $\alpha > 0$, and let Π be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^2$ with intensity $\alpha dt \otimes dz$ which will encode the sequence of times and centers of the reproduction events. We assume that the initial distribution of the population over $\mathbb{R}^2$ is encoded by $\omega_0 : \mathbb{R}^2 \to \{0, 1\}$. For all $(t, z) \in \Pi$, if the area empty at time $t-$ is encoded by $\omega_{t-} : \mathbb{R}^2 \to \{0, 1\}$, we distinguish two cases:

1. If the ball $\mathcal{B}_{\mathcal{R}}(z)$ with center z and radius $\mathcal{R}$ is Lebesgue-everywhere empty, that is, if

$$\int_{\mathcal{B}_{\mathcal{R}}(z)} \omega_{t-}(z')dz' = \mathrm{Vol}(\mathcal{B}_{\mathcal{R}}(z)),$$

   then for all $z' \in \mathcal{B}_{\mathcal{R}}(z)$, we set $\omega_t(z') = 1$. In other words, there is no real individual to reproduce in $\mathcal{B}_{\mathcal{R}}(z)$ and just after the reproduction event, the ball $\mathcal{B}_{\mathcal{R}}(z)$ stays Lebesgue-everywhere empty.

2. Otherwise, for all $z' \in \mathcal{B}_{\mathcal{R}}(z)$, we set $\omega_t(z') = 0$, and the whole ball $\mathcal{B}_{\mathcal{R}}(z)$ becomes occupied with real individuals.

The density ω_t is left unchanged outside the ball $\mathcal{B}_{\mathcal{R}}(z)$.

Again, informally, $(t, z) \in \Pi$ corresponds to a reproduction event occurring in the ball $\mathcal{B}_{\mathcal{R}}(z)$ at time t. If the corresponding area is empty, then it stays empty. But if it contains some real individuals, then one of them reproduces and completely fills the area with its descendants. When we deal with genealogies, we will assume that the reproducing individual is chosen uniformly at random among the real individuals occupying the ball just before the event (though we will actually not need to specify an ancestor in our approach below).

Since an infinite number of reproduction events occur in any time interval over the whole $\mathbb{R}^2$, this definition - although very visual - is only informal. Rigorously, the ∞-parent SLFV is a measure-valued process, defined in [Lou21] as the limit of a sequence of k-parent SLFVs when $k \to +\infty$. This process is solution to a martingale problem, which has a unique solution when reproduction events are bounded or under some condition on the distribution of the shape of reproduction events (see [Lou21]). See Section 4.2 for a rigorous construction of the ∞-parent SLFV when reproduction events affect ellipses.

In this article, we are interested in the growth properties of the occupied area in the ∞-parent SLFV and we focus on the following questions: Is the growth linear in time ? What is the speed of growth of the process ? How is it affected by the shape of the reproduction events ? In particular, we would like to compare the growth properties of the occupied region in the ∞-parent SLFV to the ones of the Eden growth model, and check that the ∞-parent SLFV has indeed similar properties. In order to do so, we first give a brief overview of what is known of the growth properties of the Eden model. This model belongs to a wider family of stochastic growth models on a lattice (generally $\mathbb{Z}^d$, $d \geq 2$) known as *first-passage percolation models*. In these models, each vertex is either occupied, or empty. Moreover, if the vertex $x \in \mathbb{Z}^d$ becomes occupied at time t, and if it is connected to a vertex $y \in \mathbb{Z}^d$ by an edge (denoted e), then vertex y becomes occupied at time $t + \tau_e$, where τ_e is independent from one edge to another (but not necessarily identically distributed). Whether the expansion is linear in time depends on the distribution of the time needed to pass through any given edge of the grid (see e.g [ADH17]) and whether one considers short-range percolation, such as nearest neighbour percolation, or long-range percolation, in which two vertices $x, y \in \mathbb{Z}^d$ are connected by an edge no matter the distance between them [CD16; CD81; Ric73]. In particular, when edge passing times are distributed as in the Eden growth model, growth is linear in time for

"short-range" percolation and potentially faster for long-range percolation, depending on the relation between the distribution of τ_e and the distance between the two vertices it connects [ADH17; CD16].

When the growth is linear in time, in general it is possible to obtain lower and upper bounds on the speed of growth (see e.g [AP02; BK93]), or to use simulations to approximate it [AD15]. For other growth models, such as the corner growth model [Sep09], which belongs to the family of *last-passage percolation models*, it is possible to obtain an explicit speed of growth for specific passage time distributions [Ros81]. As many other growth models, the Eden model is conjectured to belong to the universality class of the Kardar-Parisi-Zhang (KPZ) equation [KPZ86]. This equation generates rough fronts, whose characteristics are similar to the ones of fronts observed in some expanding biological populations (see e.g [Hue+10]). Such a conjecture is notably difficult to establish; to our knowledge, it has only been demonstrated in the case of the solid-on-solid (SOS) growth model in [BG97].

Regarding the ∞-parent SLFV, since reproduction events affect small areas, we use a comparison with a short-range percolation model to show that the growth of the occupied region is at most linear in time. Combining this with sub-additivity arguments, we obtain that the growth is linear in time, just as for the Eden model. We also obtain a lower bound on the speed of growth, and use numerical simulations to obtain a better approximation, which turns out to be significantly higher than the one initially conjectured. The simulations, presented in Section 4.2.4, suggest that the growth of the process is driven by "spikes" which occur at the front and then thicken in all directions. In order to understand how the spikes phenomenon can make the front advance faster, we consider a simple toy model in Section 4.5, composed of two interacting growing piles of cubes. We are able to obtain an explicit expression for the speed of growth of this process. The results on the toy model are in line with the numerical observations, and are also of interest in their own right.

From a population genetics viewpoint, the ∞-parent SLFV can be seen as modeling the spread of an extremely advantageous gene, corresponding to the "real" type. For weaker strengths of selection, the most classical tool to study this question is the Fisher-KPP equation [Fis37; KPP37], defined as follows. If the proportion of individuals of the favoured type at the spatial location $x \in \mathbb{R}^d$, $d \geq 1$, and at time $t \geq 0$ is given by $p(t,x) \in [0,1]$, then $p(t,x)$ evolves according to the Fisher-KPP equation if it is a (weak) solution of the equation

$$\frac{\partial p}{\partial t}(t,x) = \frac{m}{2}\Delta p(t,x)dt + s_0 p(t,x)(1-p(t,x)), \quad \forall x \in \mathbb{R}^d, \forall t \geq 0, \tag{4.1.1}$$

where $m \geq 0$ and $s_0 \geq 0$ are respectively the diffusion and selection parameters.

In dimension 1, it is possible to add stochasticity through a Wright-Fisher noise term, though this cannot be done in higher dimensions. Equation (4.1.1) and its stochastic counterpart in dimension 1 both admit traveling wave solutions (see e.g [MS95]), corresponding to a linear speed of spread/-growth. However, the spatio-temporal scales at which the spread of a weakly advantageous allele is modeled by the Fisher-KPP equation are very different from the ones in which the spread of an extremely advantageous gene is visible. The approaches and results based on the model developed here and on the Fisher-KPP equation are thus difficult to compare.

The paper is structured as follows. In Section 4.2, we define the variant of the ∞-parent SLFV with elliptical reproduction events rigorously, introduce its dual, and define what we mean by the "speed of growth" of the occupied region in the process. We also state the main result of the paper, which corresponds to Theorem 4.2.11, and analyse numerical simulations of the ∞-parent SLFV in Section 4.2.4.

The proof of Theorem 4.2.11 spans two sections. In Section 4.3, we show that the growth is *at least* linear in time, and provide a lower bound on the speed of growth of the process. In Section 4.4, we use a comparison with a first-passage percolation process to show that the growth is *at most* linear in time. Combining the results from both sections yield Theorem 4.2.11.

In Section 4.5, we study the toy model of interacting growing piles of cubes, and obtain an explicit expression for its speed of growth. This result is of interest in its own right, but also explains the discrepancy between the numerical approximation of the speed of growth of the occupied area in the ∞-parent SLFV and the value initially conjectured.

4.2 The ∞-parent spatial Lambda-Fleming Viot process with elliptical reproduction events

In this section, we rigorously define the ∞-parent SLFV, in the version we use in this article. We recall that the process was originally defined with reproduction event occuring in balls, while here we consider that they rather occur in ellipses, as it is straightforward to generalize the construction in [Lou21] to our case. Then, we introduce the dual process of potential ancestors associated to the ∞-parent SLFV. We conclude by formalizing what we mean by the speed of growth of the occupied area, and by explaining how to study it using the dual process.

4.2.1 Definition of the process

All the random objects we consider in this section are defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We use the notation $\mathbb{N} = \{0, 1, 2, ...\}$. In all that follows, let $\tilde{\mu}$ be a finite measure on $(0, +\infty)^2 \times (-\pi/2, \pi/2)$ such that there exists $\mathcal{R} > 0$ satisfying

$$\tilde{\mu}\left(\Big((0,\mathcal{R}]\times(0,\mathcal{R}]\Big)^c \times (-\pi/2,\pi/2)\right) = 0.$$

Let $\mathcal{R}_{\tilde{\mu}}$ be the smallest $\mathcal{R} > 0$ such that this condition is satisfied. We also set

$$S_{\tilde{\mu}} = (0, \mathcal{R}_{\tilde{\mu}}] \times (0, \mathcal{R}_{\tilde{\mu}}] \times (-\pi/2, \pi/2).$$

Ellipses We first set the notation regarding ellipses.

Definition 4.2.1. *Let $z_c = (x_c, y_c) \in \mathbb{R}^2$, $(a, b) \in (0, +\infty)^2$ and $\gamma \in (-\pi/2, \pi/2)$. The ellipse with center z_c and parameters (a, b, γ), denoted by $\mathfrak{B}_{a,b,\gamma}(z_c)$, is defined by:*

$$\mathfrak{B}_{a,b,\gamma}(z_c) = \left\{ \begin{pmatrix} x_c \\ y_c \end{pmatrix} + A_\gamma \begin{pmatrix} ar\cos(\theta) \\ br\sin(\theta) \end{pmatrix} : r \in [0,1], \theta \in [0, 2\pi) \right\}$$

where

$$A_\gamma = \begin{pmatrix} \cos(\gamma) & -\sin(\gamma) \\ \sin(\gamma) & \cos(\gamma) \end{pmatrix}.$$

See Figure 4.1 for an illustration. We denote the volume of an ellipse with parameters (a, b, γ) by $V_{a,b,\gamma} := \mathsf{Vol}(\mathfrak{B}_{a,b,\gamma}(0))$.

For all $(a, b) \in (0, +\infty)^2$ and $\gamma \in (-\pi/2, \pi/2)$, if f is an element of the space $C_c(\mathbb{R}^2)$ of all continuous and compactly supported functions $\mathbb{R}^2 \to \mathbb{R}$, let

$$\mathsf{Supp}^{a,b,\gamma}(f) := \left\{ z \in \mathbb{R}^2 : \mathsf{Vol}\left(\mathfrak{B}_{a,b,\gamma}(z) \cap \mathsf{Supp}(f)\right) \neq 0 \right\},$$

where $\mathsf{Supp}(f)$ stands for the support of f. The set $\mathsf{Supp}^{a,b,\gamma}(f)$ can be interpreted as the set of all potential centres $z \in \mathbb{R}^2$ for ellipses with parameters (a, b, γ) overlapping the support of f. For all $z \in \mathbb{R}^2$ and $\omega : \mathbb{R}^2 \to \{0, 1\}$ Lebesgue-measurable, let $\Theta_z^{a,b,\gamma}(\omega) : \mathbb{R}^2 \to \{0, 1\}$ be the function defined by

$$\Theta_z^{a,b,\gamma}(\omega) := \mathbb{1}_{\{\mathfrak{B}_{a,b,\gamma}(z)^c\}} \times \omega.$$

If ω represents the density of type 0 (or ghost) individuals, then $\Theta_z^{a,b,\gamma}(\omega)$ corresponds to filling the ellipse $\mathfrak{B}_{a,b,\gamma}(z)$ with type 1 individuals, without affecting the rest of $\mathbb{R}^2$.

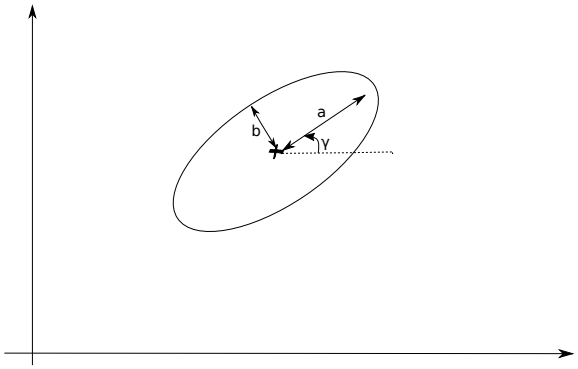


Figure 4.1: Ellipse with parameters (a, b, γ).

State space We now introduce the state space over which the ∞-parent SLFV is defined. Let $\widetilde{\mathcal{M}}_\lambda$ be the space of all measures on $\mathbb{R}^2 \times \{0,1\}$ whose marginal distribution over $\mathbb{R}^2$ is Lebesgue measure. In other words, $\widetilde{\mathcal{M}}_\lambda$ is the space of all measures M on $\mathbb{R}^2 \times \{0,1\}$ such that the following property is satisfied:

$$\forall f \in C_c(\mathbb{R}^2), \int_{\mathbb{R}^2 \times \{0,1\}} f(z) M(dz, dk) = \int_{\mathbb{R}^2} f(z) dz.$$

By a standard decomposition theorem , for all $M \in \widetilde{\mathcal{M}}_\lambda$, there exists $\omega : \mathbb{R}^2 \to [0,1]$ measurable such that

$$M(dz, dk) = (\omega(z)\delta_0(dk) + (1 - \omega(z))\delta_1(dk))dx. \tag{4.2.1}$$

Let $\mathcal{M}_\lambda$ be the set of all measures $M \in \widetilde{\mathcal{M}}_\lambda$ such that there exists $\omega : \mathbb{R}^2 \to \{0,1\}$ (instead of $[0,1]$) satisfying Eq. (4.2.1). For all $M \in \mathcal{M}_\lambda$, we refer to any measurable function $\omega : \mathbb{R}^2 \to \{0,1\}$ satisfying Eq. (4.2.1) as a *density* of M, and denote it by ω_M. Note that ω_M is not unique, but two densities will only differ at a Lebesgue-null set of points. Therefore, since we will only consider integrals of continuous functions with respect to the measures describing the current state of the population, the choice of the density ω_M used in the analysis below will not matter.

We endow $\mathcal{M}_\lambda$ with the topology of vague convergence. Let $D_{\mathcal{M}_\lambda}[0, +\infty)$ be the space of all càdlàg $\mathcal{M}_\lambda$-valued paths, endowed with the standard Skorokhod topology.

Remark 4.2.2. Assume that the state of the population at time t is encoded by $M \in \mathcal{M}_\lambda$. Since the density ω_M is not uniquely defined, for any given location $z \in \mathbb{R}^2$, $\omega_M(z)$ is not uniquely defined. Hence, it cannot be interpreted as the type of the individuals living there. Note that however, in Definition 4.2.8, we will introduce a way to assign uniquely a type to any given location, which does not depend on the choice of the density.

Martingale problem When reproduction events have bounded shape parameters, it is possible to define the ∞-parent SLFV as the solution to a well-posed martingale problem. In order to do so, we need to introduce some more notation. Let $C^1(\mathbb{R})$ be the space of all continuously differentiable functions $F : \mathbb{R} \to \mathbb{R}$. For all $f \in C_c(\mathbb{R}^2)$ and $F \in C^1(\mathbb{R})$, if $\omega : \mathbb{R}^2 \to \{0,1\}$ is measurable, we set

$$\langle \omega, f \rangle := \int_{\mathbb{R}^2} f(z)\omega(z)dz.$$

Moreover, we define the function $\Psi_{F,f} : M \in \mathcal{M}_\lambda \to \Psi_{F,f}(M) \in \mathbb{R}$ by

$$\forall M \in \mathcal{M}_\lambda, \Psi_{F,f}(M) := F\left(\langle \omega_M, f \rangle\right).$$

The functions $\Psi_{F,f}$ with $F \in C^1(\mathbb{R})$ and $f \in C_c(\mathbb{R}^2)$ will be used as the set of functions on which the martingale problem characterizing the process of interest is defined. This martingale problem is associated to the operator $\mathcal{L}^{\infty}_{\tilde{\mu}}$ defined as follows. For all $M \in \mathcal{M}_\lambda$,

$$\mathcal{L}^{\infty}_{\tilde{\mu}}\Psi_{F,f}(M) := \int_{S_{\tilde{\mu}}} \int_{\mathsf{Supp}^{a,b,\gamma}(f)} \times \left(1 - \delta_0\left(\int_{\mathfrak{B}_{a,b,\gamma}(z)} \left(1 - \omega_M(z')\right) dz'\right)\right) \times \left(F\left(\langle \Theta^{a,b,\gamma}_z(\omega_M), f\rangle\right) - F\left(\langle \omega_M, f\rangle\right)\right) dz\tilde{\mu}(da, db, d\gamma).$$

This operator encodes exactly the dynamics described informally in the introduction. Indeed, whenever a reproduction event occur, if $z \in \mathbb{R}^2$ and $(a, b, \gamma) \in S_{\tilde{\mu}}$ are the center and parameters of the corresponding ellipse, then:

1. If the affected area $\mathfrak{B}_{a,b,\gamma}(z)$ contains a positive fraction of (real) type 1 individuals (i.e, if $\int_{\mathfrak{B}_{a,b,\gamma}(z)} (1 - \omega_M(z'))\, dz' \neq 0$), then it is filled with type 1 individuals, as encoded by the action of $\Theta^{a,b,\gamma}_z$ on ω_M.

2. Otherwise, it stays filled with (ghost) type 0 individuals.

We then have the following characterization of the ∞-parent SLFV process with elliptical reproduction events.

Theorem 4.2.3. *For all $M^0 \in \mathcal{M}_\lambda$, there exists a unique $D_{\mathcal{M}_\lambda}[0, +\infty)$-valued process $(M_t)_{t\geq 0}$ such that $M_0 = M^0$ and, for all $F \in C^1(\mathbb{R})$ and $f \in C_c(\mathbb{R}^2)$,*

$$\left(\Psi_{F,f}(M_t) - \Psi_{F,f}(M_0) - \int_0^t \mathcal{L}^{\infty}_{\tilde{\mu}}\Psi_{F,f}(M_s)ds\right)_{t\geq 0}$$

is a martingale.

Moreover, this process is Markovian.

The proof of this theorem is a direct generalization of the proofs of the second part of Theorem 10 and of Lemma 21 in [Lou21], and so we omit it.

Definition 4.2.4. *Let $M^0 \in \mathcal{M}_\lambda$. The ∞-parent spatial Λ-Fleming Viot process with elliptical reproduction events (or ∞-parent SLFV) with initial condition M^0 associated to $\tilde{\mu}$ is the unique solution to the martingale problem $(\mathcal{L}^{\infty}_{\tilde{\mu}}, \delta_{M^0})$ stated in Theorem 4.2.3.*

The interest of the ∞-parent SLFV as a model for expanding populations lies in the fact that it is associated to a dual process of *potential ancestors*, which can be used to study the properties of the ∞-parent SLFV. We now define this dual process, and state the duality relation.

4.2.2 Dual process and duality relation

The dual process introduced in this section is defined on a different probability space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \boldsymbol{P})$.

State space We first introduce the state space over which the dual process is defined. Let $\mathcal{E}^c$ be the set of all subsets of $\mathbb{R}^2$ which are Lebesgue-measurable, connected, and whose Lebesgue measure is finite and non-zero. The state space we consider is then the set $\mathcal{E}^{cf}$ of all finite unions of elements of $\mathcal{E}^c$.

Definition of the dual process We can now define the dual process, called the ∞-parent ancestral process.

Definition 4.2.5. *Let $\overleftarrow{\Pi}$ be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^2 \times S_{\tilde{\mu}}$ with intensity $dt \otimes dz \otimes \tilde{\mu}(da, db, d\gamma)$, defined on the probability space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \boldsymbol{P})$, and let $E^0 \in \mathcal{E}^{cf}$. The $\mathcal{E}^{cf}$-valued ∞-parent ancestral process $(E_t^\infty)_{t\geq 0}$ with initial condition E^0 associated to $\tilde{\mu}$ is defined as follows.*

First, we set $E_0^\infty = E^0$. Then, for all $(t, z, a, b, \gamma) \in \overleftarrow{\Pi}$, if $E_{t-}^\infty \cap \mathfrak{B}_{a,b,\gamma}(z)$ has non-zero Lebesgue measure, we set

$$E_t^\infty = E_{t-}^\infty \cup \mathfrak{B}_{a,b,\gamma}(z).$$

Lemma 4.2.6. *The process $(E_t^\infty)_{t\geq 0}$ introduced in Definition 4.2.4 is well-defined and Markovian.*

Proof. We recall that for all $(t, z, a, b, \gamma) \in \overleftarrow{\Pi}$, we have $(a, b) \in (0, \mathcal{R}_{\tilde{\mu}}]^2$. Therefore,

$$\mathfrak{B}_{a,b,\gamma}(z) \subseteq \mathcal{B}_{\mathcal{R}_{\tilde{\mu}}}(z) \text{ a.s.},$$

where $\mathcal{B}_{\mathcal{R}_{\tilde{\mu}}}(z)$ is the ball of radius $\mathcal{R}_{\tilde{\mu}}$ centered at z, and so we can bound the jump rate of $(E_t^\infty)_{t\geq 0}$ from above by the one of a ∞-parent ancestral process with the same initial condition and associated to $\delta_{(\mathcal{R}_{\tilde{\mu}}, \mathcal{R}_{\tilde{\mu}}, 0)}(da, db, d\gamma)$, which is finite (see Section 3 from [Lou21]). □

Duality relation For all $M \in \mathcal{M}_\lambda$ and $E \in \mathcal{E}^{cf}$, we set

$$\widetilde{D}(M, E) := \delta_0 \left(\int_E (1 - \omega_M(z))\, dz \right).$$

Intuitively, $\widetilde{D}(M, E) = 0$ if the area E contains a positive fraction of real individuals when the population state is M, and $\widetilde{D}(M, E) = 1$ if the area is empty.

The ∞-parent SLFV and the ∞-parent ancestral process then satisfy the following duality relation, whose proof is similar to the one in [Lou21].

Proposition 4.2.7. *Let $M^0 \in \mathcal{M}_\lambda$, and let $(M_t^\infty)_{t\geq 0}$ be the unique solution to the martingale problem associated to $(\mathcal{L}_{\tilde{\mu}}^\infty, \delta_{M^0})$. Let $E^0 \in \mathcal{E}^{cf}$, and let $(E_t^\infty)_{t\geq 0}$ be the ∞-parent ancestral process with initial condition E^0 associated to $\tilde{\mu}$. Then, for all $t \geq 0$,*

$$\mathbb{E}_{M^0} \left[\widetilde{D}(M_t^\infty, E^0) \right] = \boldsymbol{E}_{E^0} \left[\widetilde{D}(M^0, E_t^\infty) \right],$$

or equivalently,

$$\mathbb{E}_{M^0} \left[\delta_0 \left(\int_{E^0} \left(1 - \omega_{M_t^\infty}(z)\right) dz \right) \right] = \boldsymbol{E}_{E^0} \left[\delta_0 \left(\int_{E_t^\infty} (1 - \omega_{M^0}(z))\, dz \right) \right].$$

This duality relation can be interpreted as follows. Whenever a reproduction event affects one area, all the individuals in the area can be considered as *potential parents*. If $\overleftarrow{\Pi}$ encodes the reproduction events affecting the population when going *backwards in time*, then E_t^∞ encodes the locations of the potential ancestors at time 0 of the individuals living in E^0 at time t. The duality relation then states that a given area E^0 contains only ghost individuals at time t if, and only if all their potential ancestors at time 0 (located in E_t^∞) are ghost individuals.

4.2.3 Speed of growth of the occupied region in the ∞-parent SLFV: Definition and main result

Let HP^0 stand for the half-plane

$$HP^0 := \{(x, y) \in \mathbb{R}^2 : x \geq 0\}.$$

In order to show that the growth of the region occupied by type 1 individuals in the ∞-parent SLFV is linear in time, we consider that initially, type 1 individuals cover the half-plane

$$\overline{HP}^0 := \left\{(x, y) \in \mathbb{R}^2 : x < 0\right\}.$$

This amounts to taking as an initial condition for the ∞-parent SLFV the measure $M^{HP}(dz) := \omega^{HP}(z)dz$, where

$$\forall z \in \mathbb{R}^2, \omega^{HP}(z) = \mathbb{1}_{z \in HP^0}.$$

Let $(M_t^{HP})_{t\geq 0}$ be the ∞-parent SLFV with initial condition M^{HP} associated to $\tilde{\mu}$. Moreover, for all $t \geq 0$, let ω_t^{HP} be a density of M_t^{HP}.

Definition 4.2.8. *For all $x \in \mathbb{R}$, let*

$$\overrightarrow{\tau}_x := \min\left\{t \geq 0 : \lim_{\epsilon\to 0} V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz < 1\right\},$$

where we recall that $\mathcal{B}_\epsilon((x, 0))$ is the ball with center $(x, 0)$ and radius ϵ, and V_ϵ is the volume of $\mathcal{B}_\epsilon((x, 0))$.

Informally, $\overrightarrow{\tau}_x$ is the first time at which the location $(x, 0)$ is reached by type 1 individuals. The fact that $\overrightarrow{\tau}_x$ can be defined as a minimum rather than an infimum is a direct consequence of the original construction of the ∞-parent SLFV introduced in Chapter 3, which implies the existence of a density $(\omega_t^\infty)_{t\geq 0}$ whose values taken over any given compact space (such that e.g. $\mathcal{B}_1((x, 0))$) are only updated at a finite rate. Moreover, as $\omega_t^{HP}(z)dz$ is measurable and $\{0, 1\}$-valued, it is also absolutely continuous with respect to Lebesgue measure, which implies that $\lim_{\epsilon\to 0} V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz$ does exist.

Remark 4.2.9. The original construction of the ∞-parent SLFV introduced in Chapter 3 also implies that in our case, at any time $t \geq 0$, the occupied area is the union of an half-plane and ellipses. This means in particular that for all $x \in \mathbb{R}$,

$$\lim_{\epsilon\to 0} V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz = 1 \quad \Longleftrightarrow \quad \exists \epsilon_0 > 0, V_{\epsilon_0}^{-1} \int_{\mathcal{B}_{\epsilon_0}((x,0))} \omega_t^{HP}(z)dz = 1.$$

Indeed, we can consider three different cases:

1. If $(x, 0)$ is out of the (closed) occupied area, then there exists ϵ_0 such that

   $$V_{\epsilon_0}^{-1} \int_{\mathcal{B}_{\epsilon_0}((x,0))} \omega_t^{HP}(z)dz = 1.$$

   Then, for all $0 < \epsilon \leq \epsilon_0$, $V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz = 1$.

2. If $(x, 0)$ is in the interior of the occupied area, then there exists ϵ_0 such that

   $$V_{\epsilon_0}^{-1} \int_{\mathcal{B}_{\epsilon_0}((x,0))} \omega_t^{HP}(z)dz = 0,$$

   which means that for all $0 < \epsilon \leq \epsilon_0$, $V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz = 0$.

3. If $(x,0)$ is on the border of the occupied area, then

$$\lim_{\epsilon\to 0} V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz \leq \frac{1}{2}.$$

Since densities are only defined up to a Lebesgue null set, we consider that $(x,0)$ is occupied by type 1 individuals if, and only if all neighborhoods of $(x,0)$ contain a non-zero fraction of type 1 individuals.

The following result means that once the location $(x,0)$ is occupied by type 1 individuals, it cannot become empty again. However, notice that the function $x \to \overrightarrow{\tau}_x$ is not necessarily increasing. See Figure 4.2 for an illustration.

Lemma 4.2.10. *Let $x \in \mathbb{R}$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left(\lim_{\epsilon\to 0} V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz < 1 \,\middle|\, \overrightarrow{\tau}_x \leq t\right) = 1.$$

Proof. We first show that for all $\epsilon > 0$,

$$t \to \mathbb{P}\left(V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz < 1\right)$$

is a non-decreasing function.

In order to do so, let $\epsilon > 0$, and let $0 \leq t \leq t'$. Then, by Proposition 4.2.7,

$$\begin{aligned}\mathbb{P}\left(V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz < 1\right) &= \mathbb{P}\left(\delta_0\left(\int_{\mathcal{B}_\epsilon((x,0))}(1-\omega_t^{HP}(z))dz\right) = 0\right)\\ &= 1 - \mathbb{E}\left[\delta_0\left(\int_{\mathcal{B}_\epsilon((x,0))}(1-\omega_t^{HP}(z))dz\right)\right]\\ &= 1 - \boldsymbol{E}\left[\delta_0\left(\int_{E_t^\infty}(1-\omega_{M^0}(z))dz\right)\right],\end{aligned}$$

where $(E_t^\infty)_{t\geq 0}$ is the ∞-parent ancestral process with initial condition $\mathcal{B}_\epsilon((x,0))$ and intensity $\tilde{\mu}$. Similarly,

$$\mathbb{P}\left(V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_{t'}^{HP}(z)dz < 1\right) = 1 - \boldsymbol{E}\left[\delta_0\left(\int_{E_{t'}^\infty}(1-\omega_{M^0}(z))dz\right)\right].$$

Moreover, $(E_t^\infty)_{t\geq 0}$ is increasing for the inclusion, so $E_t^\infty \subseteq E_{t'}^\infty$ and since ω_{M^0} is $\{0,1\}$-valued,

$$\boldsymbol{E}\left[\delta_0\left(\int_{E_t^\infty}(1-\omega_{M^0}(z))dz\right)\right] \geq \boldsymbol{E}\left[\delta_0\left(\int_{E_{t'}^\infty}(1-\omega_{M^0}(z))dz\right)\right].$$

We conclude that

$$\mathbb{P}\left(V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz < 1\right) \leq \mathbb{P}\left(V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_{t'}^{HP}(z)dz < 1\right). \tag{4.2.2}$$

Then, let $t \geq 0$, and let $(\epsilon_n)_{n\in\mathbb{N}}$ be a decreasing sequence such that $\epsilon_n \to 0$. Since

$$\lim_{\epsilon\to 0} V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz$$

exists, we have

$$\mathbb{P}\left(\lim_{\epsilon\to 0} V_\epsilon^{-1}\int_{\mathcal{B}_\epsilon((x,0))}\omega_t^{HP}(z)dz<1\middle|\overrightarrow{\tau}_x\le t\right)=\mathbb{P}\left(\lim_{n\to+\infty} V_{\epsilon_n}^{-1}\int_{\mathcal{B}_{\epsilon_n}((x,0))}\omega_t^{HP}(z)dz<1\middle|\overrightarrow{\tau}_x\le t\right).$$

Moreover, notice that the sequence $(\mathcal{B}_{\epsilon_n}((x,0)))_{n\in\mathbb{N}}$ is decreasing for the inclusion. Therefore, for all $N\in\mathbb{N}$, if $V_{\epsilon_n}^{-1}\int_{\mathcal{B}_{\epsilon_n}((x,0))}\omega_t^{HP}(z)dz=1$, then this is also the case for all $n\ge N$. Therefore,

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to+\infty} V_{\epsilon_n}^{-1}\int_{\mathcal{B}_{\epsilon_n}((x,0))}\omega_t^{HP}(z)dz=1\middle|\overrightarrow{\tau}_x\le t\right)&=\mathbb{P}\left(\exists n\in\mathbb{N}, V_{\epsilon_n}^{-1}\int_{\mathcal{B}_{\epsilon_n}((x,0))}\omega_t^{HP}(z)dz=1\middle|\overrightarrow{\tau}_x\le t\right)\\&\le\sum_{n\in\mathbb{N}}\mathbb{P}\left(V_{\epsilon_n}^{-1}\int_{\mathcal{B}_{\epsilon_n}((x,0))}\omega_t^{HP}(z)dz=1\middle|\overrightarrow{\tau}_x\le t\right)\\&\le\sum_{n\in\mathbb{N}}\mathbb{P}\left(V_{\epsilon_n}^{-1}\int_{\mathcal{B}_{\epsilon_n}((x,0))}\omega_{\overrightarrow{\tau}_x}^{HP}(z)dz=1\middle|\overrightarrow{\tau}_x\le t\right)\\&=0\end{aligned}$$

by definition of $\overrightarrow{\tau}_x$. Here we used Eq.(4.2.2) along with the fact that we condition on $\overrightarrow{\tau}_x\le t$ to pass from the second line to the third line, which allows us to conclude. □

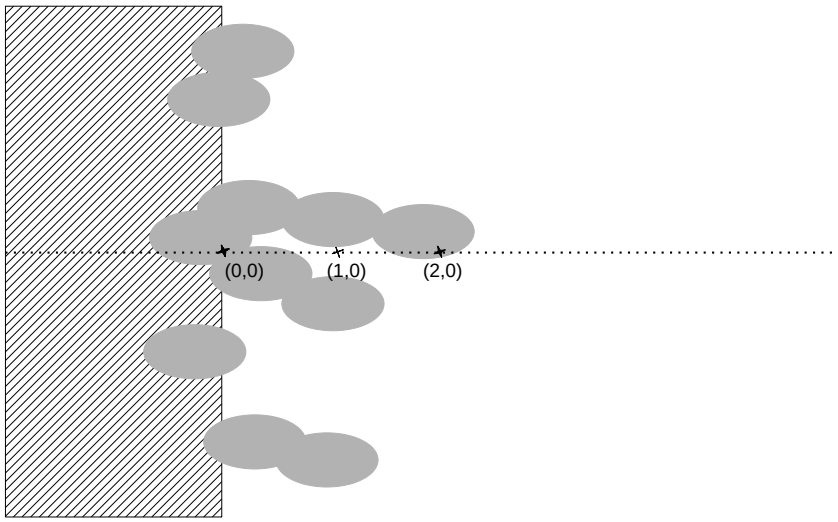


Figure 4.2: State of a population evolving according to the ∞-parent SLFV at time t. Initially, type 1 individuals cover the half-plane $\overline{HP}^0$, corresponding to the hatched area. Each grey ellipse represents a reproduction event occurring during the time interval $[0,t]$ which overlaps an area initially empty, and resulting in the corresponding area being completely filled with type 1 individuals. Here $\overrightarrow{\tau}_1>t$ but $\overrightarrow{\tau}_2\le t$.

The goal of the article is to show the following result, which tells us that the growth of the occupied area in the ∞-parent SLFV is linear in time.

Theorem 4.2.11. *There exists $\nu>0$ such that*

$$\lim_{x\to+\infty}\frac{\mathbb{E}\left[\overrightarrow{\tau}_x\right]}{x}=\nu.$$

We can then interpret ν^{-1} as the limiting speed of growth of the process. The proof can be found at the end of Section 4.4.3, preceded by a sequence of useful technical results.

In order to show Theorem 4.2.11, we make use of the dual process associated to the ∞-parent SLFV, and define an equivalent of $\overrightarrow{\tau}_x$ for the ∞-parent ancestral process. In order to do so, we first introduce a slight generalization of the ∞-parent ancestral process which allows points as initial conditions. This process is defined on the state space

$$\widetilde{\mathcal{E}}^{cf} := \mathcal{E} \cup \{\{z\} : z \in \mathbb{R}^2\}.$$

Definition 4.2.12. *Let $\overleftarrow{\Pi}$ be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^2 \times S_{\tilde{\mu}}$ with intensity $dt \otimes dz \otimes \tilde{\mu}(da, db, d\gamma)$, defined on the probability space $(\mathbf{\Omega}, \boldsymbol{\mathcal{F}}, \boldsymbol{P})$, and let $z_0 \in \mathbb{R}^2$. The $\widetilde{\mathcal{E}}^{cf}$-valued ∞-parent ancestral process $(E_t^{\infty,z_0})_{t\geq 0}$ with initial condition $\{z_0\}$ and associated to $\tilde{\mu}$ is defined as follows.*

First, we set $E_0^{\infty,z_0} = \{z_0\}$. Then, for all $(t, z, a, b, \gamma) \in \overleftarrow{\Pi}$:

- *If $E_{t-}^{\infty,z_0} = \{z_0\}$ and $z_0 \in \mathfrak{B}_{a,b,\gamma}(z)$, we set $E_t^{\infty,z_0} = \mathfrak{B}_{a,b,\gamma}$.*
- *If $E_{t-}^{\infty,z_0} \neq \{z_0\}$ and $E_{t-}^{\infty,z_0} \cap \mathfrak{B}_{a,b,\gamma}(z)$ has non-zero Lebesgue measure, we set*

$$E_t^{\infty,z_0} = E_{t-}^{\infty,z_0} \cup \mathfrak{B}_{a,b,\gamma}(z).$$

Using the same argument as for the initial ∞-parent ancestral process, $(E_t^{\infty,z_0})_{t\geq 0}$ is well-defined and Markovian. Moreover, once it jumps for the first time, its behavior is identical to the one of the original ∞-parent ancestral process.

In all that follows, let $\overleftarrow{\Pi}$ be a Poisson point process as in Definition 4.2.12. For all $x > 0$, let $(E_t^{\epsilon,x})_{t\geq 0}$, $\epsilon > 0$ be a sequence of ∞-parent ancestral processes with initial condition $\mathcal{B}_\epsilon((x, 0))$ associated to $\tilde{\mu}$, all constructed using the same underlying Poisson point process $\overleftarrow{\Pi}$. Moreover, let $(E_t^x)_{t\geq 0}$ be the ∞-parent ancestral process with initial condition $\{(x, 0)\}$ associated to $\tilde{\mu}$, also constructed using $\overleftarrow{\Pi}$. Then, $(E_t^{\epsilon,x})_{t\geq 0}$, $\epsilon > 0$ and $(E_t^x)_{t\geq 0}$ satisfy the following property.

Lemma 4.2.13. *For all $x > 0$, if t_0^x is the first time at which $(x, 0)$ is affected by a reproduction event (that is, if t_0^x is the first time at which $(E_t^x)_{t\geq 0}$ jumps), there exists almost surely $\epsilon_0^x > 0$ such that for all $t \geq t_0^x$,*

$$\forall \epsilon < \epsilon_0^x, E_t^{\epsilon,x} = E_t^x.$$

Furthermore, we have almost surely for all $t \geq 0$,

$$\lim_{\epsilon\to 0} \mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP}^0\right) = \mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right).$$

Proof. Let $x > 0$, and let $(t_0^x, z_0^x, a_0^x, b_0^x, \gamma_0^x) \in \overleftarrow{\Pi}$ be the first reproduction event to affect $(x, 0)$. In order to show the first part of the lemma, we want to show that there exists almost surely $\epsilon_0^x > 0$ such that

$$\mathcal{B}_{\epsilon_0^x}((x, 0)) \subseteq \mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x) \text{ and } \forall (t, z, a, b, \gamma) \in \overleftarrow{\Pi} \text{ such that } t < t_0^x, \mathrm{Vol}\left(\mathcal{B}_{\epsilon_0^x}((x, 0)) \cap \mathfrak{B}_{a,b,\gamma}(z)\right) = 0. \tag{4.2.3}$$

Indeed, if such a ϵ_0^x exists, then for all $0 < \epsilon < \epsilon_0^x$, the first time at which $(E_t^{\epsilon,x})_{t\geq 0}$ jumps is t_0^x. Moreover,

$$\begin{aligned} E_{t_0^x}^{\epsilon,x} &= E_0^{\epsilon,x} \cup \mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x) \\ &= \mathcal{B}_\epsilon((x, 0)) \cup \mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x) \\ &= \mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x), \\ \text{and} \quad E_{t_0^x}^x &= \mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x), \end{aligned}$$

yielding the first part of the lemma.

As the probability that x belongs to the frontier of $\mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x)$ is equal to zero, there exists almost surely $\tilde{\epsilon} > 0$ such that

$$\mathcal{B}_{\tilde{\epsilon}}((x,0)) \subseteq \mathfrak{B}_{a_0^x,b_0^x,\gamma_0^x}(z_0^x)$$

Then, we consider the number $N_{\tilde{\epsilon}}$ of reproduction events which affected $\mathcal{B}_{\tilde{\epsilon}}((x,0))$ over the time interval $[0,t_0^x)$, that is, all reproduction events $(t,z,a,b,\gamma) \in \overleftarrow{\Pi}$ such that $t \in [0,t_0^x)$ and

$$\mathcal{B}_{\tilde{\epsilon}}((x,0)) \cap \mathfrak{B}_{a,b,\gamma}(z) \neq \emptyset.$$

Since the support of $\tilde{\mu}$ is bounded, $N_{\tilde{\epsilon}}$ is almost surely finite. If $N_{\tilde{\epsilon}} = 0$, then we can set $\epsilon_0^x = \tilde{\epsilon}$ and conclude. Otherwise, let $(t_1,z_1,a_1,b_1,\gamma_1)$, ..., $(t_{N_{\tilde{\epsilon}}},z_{N_{\tilde{\epsilon}}},a_{N_{\tilde{\epsilon}}},b_{N_{\tilde{\epsilon}}},\gamma_{N_{\tilde{\epsilon}}})$ be the reproduction events which affected $\mathcal{B}_{\tilde{\epsilon}}((x,0))$ over the time interval $[0,t_0^x)$. By definition of $\tilde{\epsilon}$, for all $[\![1,N_{\tilde{\epsilon}}]\!]$, there exists $\epsilon_n > 0$ such that

$$\mathcal{B}_{\epsilon_n}((x,0)) \cap \mathfrak{B}_{a_n,b_n,\gamma_n}(z_n) \neq \emptyset,$$

and we set

$$\epsilon_0^x := \inf\{\epsilon_n : n \in [\![1,N_{\tilde{\epsilon}}]\!]\} \leq \tilde{\epsilon}.$$

As $N_{\tilde{\epsilon}}$ is almost surely finite, ϵ_0^x is almost surely strictly positive. Moreover, for all $(t,z,a,b,\gamma) \in \overrightarrow{\Pi}$ such that $t < t_0^x$, we can distinguish two cases.

1. If $\mathcal{B}_{\tilde{\epsilon}}((x,0)) \cap \mathfrak{B}_{a,b,\gamma}(z) = \emptyset$, then as $\mathcal{B}_{\epsilon_0^x}((x,0)) \subseteq \mathcal{B}_{\tilde{\epsilon}}((x,0))$,
$$\mathrm{Vol}\left(\mathcal{B}_{\epsilon_0^x}((x,0)) \cap \mathfrak{B}_{a,b,\gamma}(z)\right) = 0.$$
2. If $\mathcal{B}_{\tilde{\epsilon}}((x,0)) \cap \mathfrak{B}_{a,b,\gamma}(z) \neq \emptyset$, then $(t,z,a,b,\gamma) \in \{(t_i,z_i,a_i,b_i,\gamma_i) : i \in [\![1,N_{\tilde{\epsilon}}]\!]\}$, and by definition of ϵ_0^x,
$$\mathrm{Vol}\left(\mathcal{B}_{\epsilon_0^x}((x,0)) \cap \mathfrak{B}_{a,b,\gamma}(z)\right) = 0.$$

allowing us to conclude.

In order to show the second part of the lemma, we first consider $t \in [0,t_0^x)$, and assume that ϵ_0^x exists (which is almost surely satisfied). Then,

$$\mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right) = \mathrm{Vol}\left(\{(x,0)\} \cap \overline{HP}^0\right) = 0.$$

Moreover, by (4.2.3), for all $0 < \epsilon < \min(\epsilon_0^x, x/2)$,

$$\mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP}^0\right) = \mathrm{Vol}\left(\mathcal{B}_\epsilon((x,0)) \cap \overline{HP}^0\right) = \mathrm{Vol}(\emptyset) = 0.$$

Therefore,

$$\lim_{\epsilon\to 0} \mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP}^0\right) = 0 = \mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right).$$

Then, let $t \geq t_0^x$. For all $0 < \epsilon < \epsilon_0^x$,

$$\mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP}^0\right) = \mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right),$$
$$\text{and hence} \quad \lim_{\epsilon\to 0} \mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP}^0\right) = \mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right).$$

□

We can now introduce an equivalent of $\overrightarrow{\tau}_x$ for the ∞-parent ancestral process.

Definition 4.2.14. *For all $x > 0$, let $(t_0^x, z_0^x, a_0^x, b_0^x, \gamma_0^x) \in \overleftarrow{\Pi}$ be the first reproduction event to affect $(x, 0)$, and let*

$$\tilde{\tau}_x := \begin{cases} \min\left\{t \geq 0 : \lim_{\epsilon \to 0} \mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP^0}\right) > 0\right\} & \text{if there exists } \epsilon_0^x > 0 \text{ satisfying (4.2.3)} \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 4.2.15. *In the notation of Definition 4.2.14, for all $x > 0$, $\tilde{\tau}_x$ is well-defined, and is almost surely equal to*

$$\min\left\{t \geq 0 : E_t^x \cap \overline{HP}^0 \neq \emptyset\right\}.$$

In particular, $\tilde{\tau}_x \neq 0$ a.s.

Proof. Let $x > 0$. By Lemma 4.2.13, $\tilde{\tau}_x$ is well-defined and almost surely equal to

$$\min\{t \geq 0 : \mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right) > 0\} \geq t_0^x \text{ a.s.},$$

since the support of $\tilde{\mu}$ is bounded. Moreover, since for all $(a, b, \gamma) \in S_{\tilde{\mu}}$, the set

$$\left\{z \in \mathbb{R}^2 : \mathfrak{B}_{a,b,\gamma}(z) \cap \overline{HP}^0 \neq \emptyset \text{ and } \mathrm{Vol}\left(\mathfrak{B}_{a,b,\gamma}(z) \cap \overline{HP}^0\right) = 0\right\}$$

has zero Lebesgue measure, we have

$$\min\{t \geq 0 : \mathrm{Vol}\left(E_t^x \cap \overline{HP}^0\right) > 0\} = \min\left\{t \geq 0 : E_t^x \cap \overline{HP}^0 \neq \emptyset\right\} \text{ a.s.},$$

which allows us to conclude. □

In order to use $(\tilde{\tau}_x)_{x>0}$ to show results on $(\overrightarrow{\tau}_x)_{x>0}$, we will need the following result.

Lemma 4.2.16. *For all $x > 0$, $\tilde{\tau}_x$ and $\overrightarrow{\tau}_x$ have the same distribution.*

Proof. We recall that constructing all the ∞-parent ancestral process $(E_t^{\epsilon,x})_{t\geq 0}$ using the same underlying Poisson point process ensures that for all $t \geq 0$, for all $t \geq 0$,

$$\forall 0 < \epsilon < \epsilon', E_t^{\epsilon,x} \subseteq E_t^{\epsilon',x}. \tag{4.2.4}$$

Moreover, for all $\epsilon > 0$, we set

$$\overrightarrow{\tau}_x^\epsilon := \min\left\{t > 0 : V_\epsilon^{-1} \int_{\mathcal{B}_\epsilon((x,0))} \omega_t^{HP}(z)dz < 1\right\}$$
$$\text{and} \quad \tilde{\tau}_x^\epsilon := \min\left\{t > 0 : \mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP^0}\right) > 0\right\}.$$

Since the ancestral processes jump at a finite rate, $\tilde{\tau}_x^\epsilon$ can be defined as a minimum. As in the case of $\overrightarrow{\tau}_x$, the original definition of the ∞-parent SLFV introduced in Chapter 3 implies that it is also possible to define $\overrightarrow{\tau}_x^\epsilon$ as a minimum rather than an infimum.

Using the same argument as in the proof of Lemma 4.2.15, we obtain that

$$\tilde{\tau}_x = \lim_{\epsilon \to 0} \tilde{\tau}_x^\epsilon \quad \text{a.s.}$$

We have a similar result concerning $\overrightarrow{\tau}_x$. Indeed, for all $0 < \epsilon < \epsilon'$, by definition of $\overrightarrow{\tau}^{\epsilon}_x$,

$$\begin{aligned}\int_{\mathcal{B}_{\epsilon'}((x,0))} \omega^{HP}_{\overrightarrow{\tau}^{\epsilon}_x}(z)dz &= \int_{\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_{\overrightarrow{\tau}^{\epsilon}_x}(z)dz \\ &\quad + \int_{\mathcal{B}_{\epsilon'}((x,0))\setminus\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_{\overrightarrow{\tau}^{\epsilon}_x}(z)dz \\ &< V_\epsilon + V_{\epsilon'} - V_\epsilon \\ &< V_{\epsilon'}.\end{aligned}$$

Therefore, $\overrightarrow{\tau}^{\epsilon'}_x \leq \overrightarrow{\tau}^{\epsilon}_x$ and $\lim_{\epsilon\to 0} \overrightarrow{\tau}^{\epsilon}_x$ exists. Moreover, for all $t \geq \overrightarrow{\tau}_x$, by Lemma 4.2.10,

$$\lim_{\epsilon\to 0} V_\epsilon^{-1} \int_{\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_t(z)dz < 1,$$

so there exists $\epsilon_t > 0$ such that for all $0 < \epsilon \leq \epsilon_t$,

$$V_\epsilon^{-1} \int_{\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_t(z)dz < 1,$$

and hence for all $0 < \epsilon \leq \epsilon_t$, $t \geq \overrightarrow{\tau}^{\epsilon}_x$. Similarly, if $t > \lim_{\epsilon\to 0} \overrightarrow{\tau}^{\epsilon}_x$, then again by Lemma 4.2.10, there exists $\epsilon_t > 0$ such that for all $0 < \epsilon \leq \epsilon_t$,

$$V_\epsilon^{-1} \int_{\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_t(z)dz < 1$$

and $t > \overrightarrow{\tau}_x$. Therefore,

$$\overrightarrow{\tau}_x = \lim_{\epsilon\to 0} \overrightarrow{\tau}^{\epsilon}_x \quad \text{a.s.}$$

Thus it is sufficient to show that $\overrightarrow{\tau}^{\epsilon}_x$ and $\tilde{\tau}^{\epsilon}_x$ have the same distribution for all $\epsilon > 0$ in order to conclude the proof.

Let $\epsilon > 0$, and let $t \geq 0$. Since $t \to \int_{\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_t(z)dz$ is non-increasing and $[0, V_\epsilon]$-valued,

$$\begin{aligned}\mathbb{P}_{M_0^{HP}}\left(\overrightarrow{\tau}^{\epsilon}_x > t\right) &= \mathbb{P}_{M_0^{HP}}\left(\int_{\mathcal{B}_{\epsilon}((x,0))} \omega^{HP}_t(z)dz = V_\epsilon\right) \\ &= \mathbb{P}_{M_0^{HP}}\left(\delta_0\left(\int_{\mathcal{B}_{\epsilon}((x,0))} \left(1 - \omega^{HP}_t(z)\right) dz\right) = 1\right) \\ &= \mathbb{E}_{M_0^{HP}}\left[\delta_0\left(\int_{\mathcal{B}_{\epsilon}((x,0))} \left(1 - \omega^{HP}_t(z)\right) dz\right)\right] \\ &= \boldsymbol{E}_{\mathcal{B}_{\epsilon}((x,0))}\left[\delta_0\left(\int_{E^{\epsilon,x}_t} \left(1 - \omega^{HP}_0(z)\right) dz\right)\right] \\ &= \boldsymbol{P}_{\mathcal{B}_{\epsilon}((x,0))}\left(\delta_0\left(\int_{E^{\epsilon,x}_t} \left(1 - \omega^{HP}_0(z)\right) dz\right) = 1\right) \\ &= \boldsymbol{P}_{\mathcal{B}_{\epsilon}((x,0))}\left(\mathrm{Vol}\left(E^{\epsilon,x}_t \cap \overline{HP^0}\right) = 0\right) \\ &= \boldsymbol{P}_{\mathcal{B}_{\epsilon}((x,0))}\left(\tilde{\tau}^{\epsilon}_x > t\right).\end{aligned}$$

Here we used Proposition 4.2.7 to pass from the third to the fourth line. □

For all $x > 0$, let HP^x stand for the half-plane

$$HP^x := \left\{(x', y) \in \mathbb{R}^2 : x' \geq x\right\},$$

extending the notation HP^0 to the case $x \geq 0$.

In practice, instead of using $(\tilde{\tau}_x)_{x>0}$, it will be more convenient to work with another random variable. Indeed, if $x < x'$, even if $\tilde{\tau}_x$ and $\tilde{\tau}_{x'}$ are constructed using the same underlying Poisson point process, this is not sufficient to have $\tilde{\tau}_x \leq \tilde{\tau}_{x'}$ a.s., as the underlying ∞-parent ancestral processes have different starting positions. Therefore, we define another sequence $(\overleftarrow{\tau}_x)_{x>0}$ such that for all $x > 0$, $\overleftarrow{\tau}_x$ and $\tilde{\tau}_x$ have the same distribution, but also such that all underlying ∞-parent ancestral processes start from the same location, ensuring that for all $x < x'$, $\overleftarrow{\tau}_x \geq \overleftarrow{\tau}_{x'}$. Moreover, for all $x > 0$, the ∞-parent ancestral process associated to $\overleftarrow{\tau}_x$ can be seen as the symmetric of the one associated to $\tilde{x}$ with respect to the axis $\{(x/2, y) : y \in \mathbb{R}\}$.

Let $\tilde{\mu}^{\leftarrow}$ be a finite measure on $\mathbb{R}_+ \times \mathbb{R}^2 \times S_{\tilde{\mu}}$ such that for all I, J intervals of $(0, \mathcal{R}_{\tilde{\mu}}]$ and for all $-\pi/2 < \gamma_1 \leq \gamma_2 < \pi/2$,

$$\tilde{\mu}^{\leftarrow}(I \times J \times [\gamma_1, \gamma_2]) = \tilde{\mu}(I \times J \times [\pi - \gamma_2, \pi - \gamma_1]).$$

Let $(E_t^\epsilon)_{t\geq 0}$, $\epsilon > 0$ be a sequence of ∞-parent ancestral processes with initial condition $\mathcal{B}_\epsilon((0,0))$ associated to $\tilde{\mu}^{\leftarrow}$ constructed using the same underlying Poisson point process, and let $(E_t)_{t\geq 0}$ be the ∞-parent ancestral process with initial condition $\{(0,0)\}$ associated to $\tilde{\mu}^{\leftarrow}$, constructed using the same underlying Poisson point process as $(E_t^\epsilon)_{t\geq 0}$, $\epsilon > 0$.

Definition 4.2.17. *For all $x > 0$, let $(t_0, z_0, a_0, b_0, \gamma_0) \in \overleftarrow{\Pi}$ be the first reproduction event to affect $(0,0)$, and let*

$$\overleftarrow{\tau}_x := \begin{cases} \min\left\{t \geq 0 : \lim_{\epsilon\to 0} \mathrm{Vol}\left(E_t^\epsilon \cap HP^x\right) > 0\right\} & \text{if there exists } \epsilon_0 > 0 \\ & \text{such that } \mathcal{B}_{\epsilon_0}((0,0)) \subseteq \mathfrak{B}_{a_0,b_0,\gamma_0}(z_0), \\ 0 & \text{otherwise.} \end{cases}$$

As for $(\tilde{\tau}_x)_{x>0}$, for all $x > 0$, $\overleftarrow{\tau}_x$ is well-defined and almost surely equal to

$$\min\left\{t \geq 0 : E_t \cap HP^x \neq \emptyset\right\}.$$

If we say that a point $z = (x, y) \in \mathbb{R}^2$ is at *horizontal separation* d of the point $z' = (x', y') \in \mathbb{R}^2$ if, and only if $x - x' = d$, then informally, $\overleftarrow{\tau}_x$ represents the first time the ∞-parent ancestral process starting from $(0,0)$ reaches points at horizontal separation of at least x from the starting location. Moreover, we have the following lemma.

Lemma 4.2.18. *The function $x \to \overleftarrow{\tau}_x$ is nondecreasing.*

Proof. We distinguish two cases. If there does not exist $\epsilon_0 > 0$ satisfying (4.2.3), then for all $x > 0$, $\overleftarrow{\tau}_x = 0$, and we can conclude. Now, we assure that there exists $\epsilon_0 > 0$ satisfying (4.2.3), and let $0 < x_1 < x_2$. Then, by Lemma 4.2.13 and its proof,

$$\forall 0 < \epsilon \leq \min(\epsilon_0, x_1/2), \mathrm{Vol}\left(E_t^\epsilon \cap HP^{x_1}\right) = \mathrm{Vol}\left(E_t^{\min(\epsilon_0, x_1/2)} \cap HP^{x_1}\right)$$

and $$\forall 0 < \epsilon \leq \min(\epsilon_0, x_2/2), \mathrm{Vol}\left(E_t^\epsilon \cap HP^{x_2}\right) = \mathrm{Vol}\left(E_t^{\min(\epsilon_0, x_2/2)} \cap HP^{x_2}\right).$$

We set $\epsilon' = \min(\epsilon_0, x_1/2, x_2/2)$. Then, $HP^{x_2} = HP^{x_1}$ and so

$$\mathrm{Vol}\left(E^{\epsilon'}_{\overleftarrow{\tau}_{x_2}} \cap HP^{x_1}\right) \geq \mathrm{Vol}\left(E^{\epsilon'}_{\overleftarrow{\tau}_{x_2}} \cap HP^{x_2}\right) = \lim_{\epsilon\to 0} \mathrm{Vol}\left(E_t^\epsilon \cap HP^{x_2}\right) > 0$$

by definition of $\overleftarrow{\tau}_{x_2}$, so $\overleftarrow{\tau}_{x_1} \leq \overleftarrow{\tau}_{x_2}$ and we can conclude. □

Notice that we are now studying the expansion of the backwards-in-time process in the same direction as the occupied area in the forwards-in-time process. Conversely, $(\tilde{\tau}_x)_{x>0}$ corresponds to the expansion of the ∞-parent ancestral process in the *opposite direction*. Moreover, we recall that $(\overleftarrow{\tau}_x)_{x>0}$ can be seen as constructed using an underlying ∞-parent ancestral process which is the symmetric of the one used to construct $(\tilde{\tau}_x)_{x>0}$ with respect to the axis $\{(x/2, y) : y \in \mathbb{R}\}$. This observation yields the following lemma.

Lemma 4.2.19. *For all $x > 0$, $\overleftarrow{\tau}_x$ and $\tilde{\tau}_x$ have the same distribution.*

Proof. Let $x > 0$. For all $t \geq 0$ and $\epsilon > 0$, let $\mathrm{Sym}(E_t^{\epsilon,x})$ be the symmetric of $E_t^{\epsilon,x}$ with respect to the axis $\{(x/2, y) : y \in \mathbb{R}\}$. Then, $(\mathrm{Sym}(E_t^{\epsilon,x}))_{t\geq 0}$, $\epsilon > 0$ is a sequence of ∞-parent ancestral processes with initial condition $\mathcal{B}_\epsilon((0,0))$ and parameters $(a, b, \pi - \gamma)$ all constructed using the same Poisson point process, which can be used to construct $\overleftarrow{\tau}_x$. Moreover, for all $t \geq 0$ and $\epsilon > 0$,

$$\mathrm{Vol}\left(E_t^{\epsilon,x} \cap \overline{HP^0}\right) > 0 \quad \text{if, and only if} \quad \mathrm{Vol}\left(\mathrm{Sym}(E_t^{\epsilon,x}) \cap HP^x\right) > 0,$$

which allows us to conclude. □

In order to show Theorem 4.2.11, we will use the following proposition, which is a direct consequence of Lemmas 4.2.16 and 4.2.19.

Proposition 4.2.20. *For all $x > 0$, $\overleftarrow{\tau}_x$ and $\overrightarrow{\tau}_x$ have the same distribution.*

4.2.4 Numerical simulations

In order to obtain an approximation for the limiting speed of growth ν^{-1} by means of numerical simulations, we can use the fact that

$$\nu^{-1} = \left(\lim_{x\to+\infty} x^{-1}\mathbb{E}[\overleftarrow{\tau}_x]\right)^{-1},$$

which is a consequence of Theorem 4.2.11 combined with Proposition 4.2.20. Indeed, the ∞-parent ancestral is easier to simulate than its forwards-in-time counterpart, since it jumps at a finite rate at any time, and since there are no border effects to take into account while simulating the process on an appropriately chosen compact subset of $\mathbb{R}^2$.

We focus on the case in which all ellipses have the same shape parameters. In order to be able to compare the speed of growth of the occupied regions in ∞-parent SLFV with different shape parameters, we assume that any given location $z \in \mathbb{R}^2$ is affected by a reproduction event at rate 1. Therefore, we take

$$\mu(da, db, d\gamma) = V_{a,b,\gamma}^{-1}\delta_{a_0}(da) \otimes \delta_{b_0}(db) \otimes \delta_0(d\gamma),$$

where $a_0 \in (0, +\infty)$ and where b_0 is chosen such that the volume of the corresponding ellipse is equal to π.

For 9 different values of a ranging from 0.33 to 3, we simulate 30 ∞-parent ancestral processes with initial condition $\{(0,0)\}$ and parameters $(a, b, 0)$, where b is chosen as stated above. This ensures that the we compare ∞-parent SLFV processes for which reproduction events have the same scale.

The results can be found in Figure 4.3. The numerical simulations show that the speed of growth is a linear function of a (over the range of a-values considered). This speed is around 2.6 times higher than the lower bound obtained in Section 4.3 (equal to a when $\gamma = 0$), which was initially conjectured to be the limiting speed of the process.

Numerical simulations suggest that the growth of the process seems to be driven by "spikes" in the expansion direction, that then thicken and grow sideways, bridging the gap with the rest of the population. See Figure 4.4 for an illustration of this phenomenon. This motivates the study of a toy model of two interacting growing piles of cubes in Section 4.5.

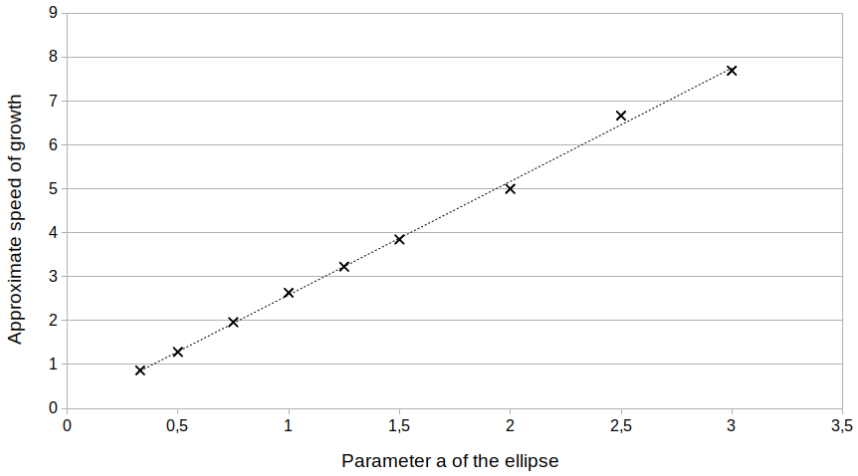


Figure 4.3: Approximate speed of growth of the occupied area in the ∞-parent SLFV process, as a function of a. For each value of a, 30 ∞-parent ancestral processes with parameters $(a, b, 0)$ were simulated, in order to compute $\mathbb{E}[\overleftarrow{\tau}_x]$ for large values of x. The crosses indicate the approximate values, and the dotted line corresponds to the line of equation $\nu^{-1} = 2.58\, a$.

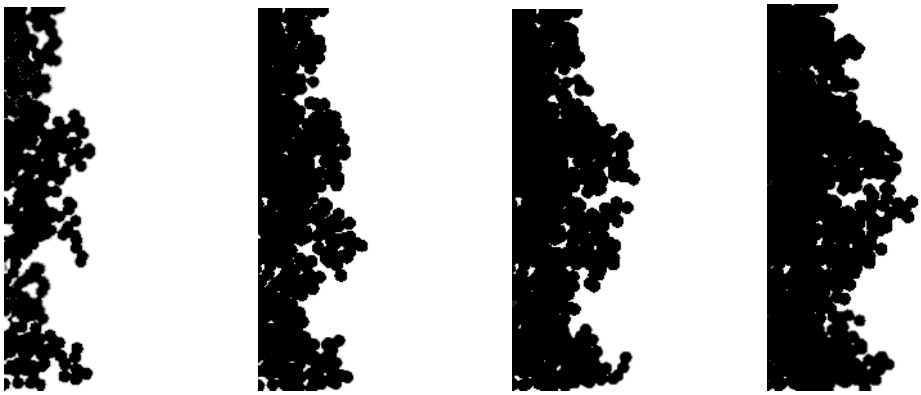


Figure 4.4: Illustration of the growth dynamics of the occupied region in the ∞-parent SLFV, with $(a, b, \gamma) = (1, 1, 0)$. The images represent the same ∞-parent SLFV at four instants $0 < t_1 < t_2 < t_3 < t_4$. The black area represents the area occupied by real individuals, and the white area is empty (or equivalently, filled with ghost individuals). The expansion starts from the left-most part of the image, and goes towards the right-most part of the image.

4.3 Lower bound on the speed of growth

In this section, we show that the growth of the occupied area in the ∞-parent SLFV, as well as the one of its dual, is *at least* linear in time. More precisely, we show the following result.

Proposition 4.3.1. *There exists $\nu \geq 0$ such that*

$$\lim_{x \to +\infty} \frac{\mathbb{E}\left[\overleftarrow{\tau}_x\right]}{x} = \nu.$$

The proof of this result can be found at the end of Section 4.3.3. By Proposition 4.2.20, this result is then also true for $\overrightarrow{\tau}_x$.

The key difference with Theorem 4.2.11 is that this proposition does not require ν to be non-zero, and hence only means that the growth is at least linear in time (and faster if $\nu = 0$). In Section 4.4, we show that we indeed have $\nu \neq 0$, or in other words, that the growth is exactly linear in time.

Remark 4.3.2. Since the limiting speed of growth is given by ν^{-1}, a lower bound on the speed of growth amounts to an *upper bound* on $\lim_{x \to +\infty} x^{-1}\mathbb{E}\left[\overleftarrow{\tau}_x\right]$.

In order to prove Proposition 4.3.1, we first establish a few auxiliary results. In all that follows, let Π be a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}^2 \times S_{\tilde{\mu}}$ with intensity $dt \otimes dz \otimes \tilde{\mu}^{\leftarrow}$, and let $(E_t)_{t \geq 0}$ be the ∞-parent ancestral process with initial condition $\{(0,0)\}$ and associated to $\tilde{\mu}^{\leftarrow}$, constructed using the Poisson point process Π. Here we use the modification of the ∞-parent ancestral process introduced earlier, which allows to take simple points as initial conditions.

4.3.1 Sub-additivity

We first introduce other ∞-parent ancestral processes, coupled to $(E_t)_{t \geq 0}$ via the Poisson point process Π. For all $z \in \mathbb{R}^2$ and $s \geq 0$, let $(E_t^{z,s})_{t \geq 0}$ be the ∞-parent ancestral process with initial condition $\{z\}$ associated to $\tilde{\mu}^{\leftarrow}$, constructed using *only the reproduction events in Π occurring strictly after time s*. If $s = 0$, then $(E_t^{z,0})_{t \geq 0}$ is the regular ∞-parent ancestral process, but if $s \neq 0$, then the process is constant and equal to $\{z\}$ on the time interval $[0, s]$, and only after does it start following the dynamics of an ∞-parent ancestral process.

Since all the processes $(E_t^{z,s})_{t \geq 0}$, $z \in \mathbb{R}^2$, $s \geq 0$ are constructed using the same underlying Poisson point process, we have the following lemma.

Lemma 4.3.3. *For all $z_1, z_2 \in \mathbb{R}^2$ and $0 < s_1 < s_2$, if $z_2 \in E_{s_2}^{z_1,s_1}$, then for all $t \geq s_2$,*

$$E_t^{z_2,s_2} \subseteq E_t^{z_1,s_1} \quad \textit{almost surely}.$$

We now introduce the following family of random variables. First, for all $n \in \mathbb{N}$, let

$$\begin{aligned} T_{0,n} :&= \min\left\{t \geq 0 : E_t^{(0,0),0} \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \neq \emptyset\right\} \\ &= \min\left\{t \geq 0 : E_t \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \neq \emptyset\right\} \\ &= \overleftarrow{\tau}_{4n\mathcal{R}_{\tilde{\mu}}}, \end{aligned}$$

Here $T_{0,n}$ can be defined as a minimum rather than an infimum due to the fact that the ∞-parent ancestral process jumps at a finite rate. Then, let P_n be sampled uniformly at random among the points in $E_{T_{0,n}}$ at horizontal separation of *exactly* $4n\mathcal{R}_{\tilde{\mu}}$ from $(0,0)$. That is, P_n is a uniform sample from the compact set

$$E_{T_{0,n}} \cap \left\{(x,y) \in \mathbb{R}^2 : x = 4n\mathcal{R}_{\tilde{\mu}}\right\}.$$

Moreover, for all $0 \leq m \leq n \in \mathbb{N}$, let

$$T_{m,n} := \min\left\{t \geq 0 : E_{t+T_{0,m}}^{P_m,T_{0,m}} \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \neq \emptyset\right\},$$

where $E_{t+T_{0,m}}^{P_m,T_{0,m}}$ corresponds to the ∞-parent ancestral process started from P_m at time $T_{0,m}$.

By construction, the family $(T_{m,n})_{0\leq m\leq n}$ satisfies the following lemma.

Lemma 4.3.4. *For all* $0 < m < n \in \mathbb{N}$*,*

$$T_{0,n} \leq T_{0,m} + T_{m,n}.$$

Proof. Let $0 < m < n \in \mathbb{N}$. By definition of the random variables,

$$P_m \in E_{T_{0,m}} = E_{T_{0,m}}^{(0,0),0}.$$

Therefore, using Lemma 4.3.3, for all $t \geq 0$,

$$E_{t+T_{0,m}}^{P_m,T_{0,m}} \subseteq E_{t+T_{0,m}}.$$

In particular, this is true for $t = T_{m,n}$, hence

$$E_{T_{m,n}+T_{0,m}}^{P_m,T_{0,m}} \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \subseteq E_{T_{m,n}+T_{0,m}} \cap HP^{4n\mathcal{R}_{\tilde{\mu}}},$$

from which we deduce (by definition of $T_{0,m}$, P_m and $T_{m,n}$)

$$E_{T_{m,n}+T_{0,m}} \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \neq \emptyset.$$

Therefore,

$$\begin{aligned} T_{0,n} &= \min\left\{t > 0 : E_t \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \neq \emptyset\right\} \\ &\leq T_{m,n} + T_{0,m}. \end{aligned}$$

□

By invariance by translation of the underlying Poisson point processes, the following lemmas are also true.

Lemma 4.3.5. *For all* $n \in \mathbb{N}$*, the joint distribution of* $(T_{n+1,n+k+1})_{k\geq 1}$ *is the same as the one of* $(T_{n,n+k})_{k\geq 1}$.

Lemma 4.3.6. *For all* $k \in \mathbb{N}\backslash\{0\}$*, the random variables* $(T_{nk,(n+1)k})_{n\geq 1}$ *are i.i.d.*

In order to use Theorem 1.10 from [Lig85] and conclude, all we need is to show that the following lemma is true.

Lemma 4.3.7. *For all* $n \in \mathbb{N}$*,* $\mathbb{E}[T_{0,n}] < +\infty$.

Lemma 4.3.7 is proved at the end of Section 4.3.3, using the so-called *express chain*. We will then use it to complete the proof of Proposition 4.3.1 (again at the end of Section 4.3.3).

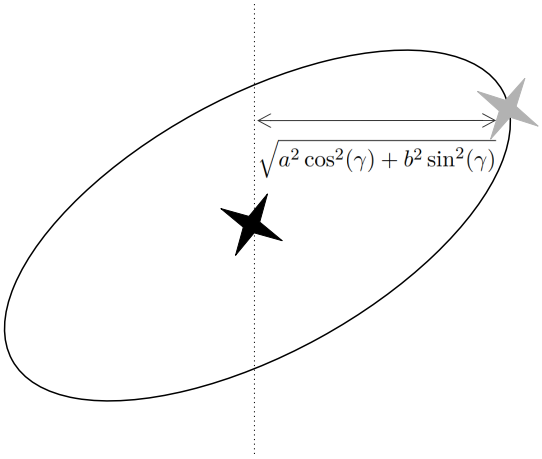


Figure 4.5: Ellipse with parameters (a, b, γ).

4.3.2 Definition of the express chain

The use of what we will call the *express chain* can be motivated by the following observation. In each ellipse with center $z = (x, y) \in \mathbb{R}^2$ and parameters (a, b, γ), there exists exactly one point for which the horizontal separation from the center is maximal: the one with coordinates

$$(x + a\cos(\theta_{max})\cos(\gamma) - b\sin(\theta_{\max})\sin(\gamma), y + a\cos(\theta_{max})\sin(\gamma) + b\sin(\theta_{max})\cos(\gamma)),$$

where $\theta_{max} = \arctan(-ba^{-1}\tan(\gamma))$. This point is at horizontal separation $\sqrt{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}$ from z (see Figure 4.5).

If we take this potential parent, wait until it is affected by a new reproduction event, and repeat, we obtain a Markov jump process, jumping at rate 1 and going away from 0 at an average *horizontal speed* of $\mathbb{E}[\sqrt{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}]$ (modulo some stochasticity due to the location of the center of the reproduction event). See Appendix 4.6 for the proof of the geometrical properties of ellipses used throughout this section.

Formally, the express chain is defined as follows.

Definition 4.3.8. *The express chain associated to $(E_t)_{t\geq 0}$ (constructed using Π), denoted $(C_t^{express})_{t\geq 0}$, is the $\mathbb{R}^2$-valued Markov process defined as follows.*

First, we set $C_0^{express} = (0, 0)$. Then, for all $(t, z_c, a, b, \gamma) \in \Pi$, if $C_{t-}^{express} \in \mathfrak{B}_{a,b,\gamma}(z_c)$ and $z_c = (x_c, y_c)$, we set:

$$C_t^{express} = (X_t^{express}, Y_t^{express})$$

where

$$\begin{aligned} X_t^{espress} &= x_c + a\cos(\theta_{max})\cos(\gamma) - b\sin(\theta_{\max})\sin(\gamma), \\ Y_t^{express} &= y_c + a\cos(\theta_{max})\sin(\gamma) + b\sin(\theta_{max})\cos(\gamma) \end{aligned}$$

and $\theta_{max} = \arctan(-ba^{-1}\tan(\gamma))$.

See Figure 4.6 for an illustration of how to construct the express chain.

The interest of the express chain lies in the following observation, whose proof is a direct consequence of the definition of $T_{0,n}$.

Lemma 4.3.9. *Let $n \in \mathbb{N}^*$. For all $t \geq 0$, if $C_t^{express} \in HP^{4n\mathcal{R}_{\tilde{\mu}}}$, then $T_{0,n} \leq t$ a.s.*

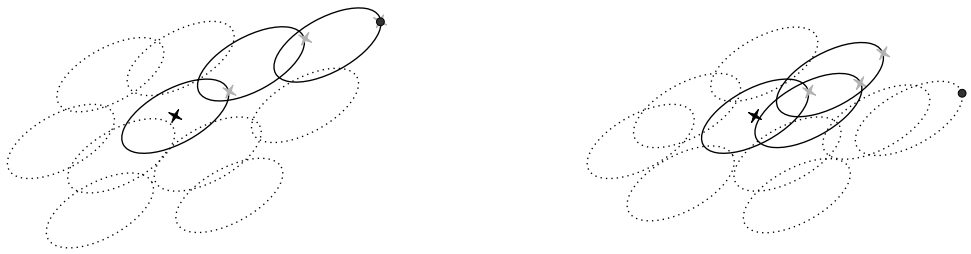


Figure 4.6: Comparison between the express chain and the ∞-parent ancestral process at time t, for two realizations of the ∞-parent ancestral process. The ellipses indicate the reproduction events which affected the ∞-parent ancestral process until time t. The crosses represent the successive positions of the express chain, and the black point indicates the location reached by the ∞-parent ancestral process at time t which maximizes the horizontal separation from the starting location. This location coincides with the one reached by the express chain in the first case, but not in the second case.

Proof. Let $t \geq 0$, and assume $C_t^{express} \in HP^{4n\mathcal{R}_{\tilde{\mu}}}$. Then,

$$E_t \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \supset \{C_t^{express}\} \cap HP^{4n\mathcal{R}_{\tilde{\mu}}} \supset \{C_t^{express}\} \neq \emptyset,$$

and so $t \geq T_{0,n}$. □

Therefore, if for all $n \in \mathbb{N}$, we set

$$T_{0,n}^{express} := \min\left\{t \geq 0 : C_t^{express} \in HP^{4n\mathcal{R}_{\tilde{\mu}}}\right\},$$

then for all $n \in \mathbb{N}$,

$$T_{0,n} \leq T_{0,n}^{express} \text{a.s.} \tag{4.3.1}$$

$$\text{and} \quad \mathbb{E}[T_{0,n}] \leq \mathbb{E}[T_{0,n}^{express}]. \tag{4.3.2}$$

In other words, in order to show Lemma 4.3.7, it is sufficient to obtain a similar result on $\mathbb{E}[T_{0,n}^{express}]$.

Before studying the properties of the express chain in the next section, we introduce the following notation. For all $t \geq 0$, let $N_t^{express}$ be the number of jumps of the express chain on the time interval $[0,t]$. For all $i \in \mathbb{N}\backslash\{0\}$, let t_i be the instant of the i-th jump of the express chain, let $R_i = (R_i^X, R_i^Y)$ be the coordinates of the center of the reproduction event triggering this jump, and let (a_i, b_i, γ_i) be the parameters of the ellipse affected by the reproduction event. We then set

$$D_i = \sqrt{a_i^2 \cos^2(\gamma_i) + b_i^2 \sin^2(\gamma_i)}.$$

In other words, for all $i \in \mathbb{N}\backslash\{0\}$, D_i is a random variable encoding to the distance between the center of the reproduction event and the right-most point in the corresponding ellipse. Notice that the random variables $(D_i)_{i\in\mathbb{N}\backslash\{0\}}$ are i.i.d.

4.3.3 Properties of the express chain

We now study some properties of the express chain, in order to obtain an upper bound on $\mathbb{E}[T_{0,n}^{express}]$.

Let $t \geq 0$. By construction,

$$\begin{aligned} X_t^{express} &= \sum_{i=1}^{N_t^{express}} \left(R_i^X - X_{t_i-}^{express} + D_i\right) \\ &= \sum_{i=1}^{N_t^{express}} \left(R_i^X - X_{t_i-}^{express} + D_i - \mathbb{E}[D_1]\right) + \mathbb{E}[D_1] \\ &= \sum_{i=1}^{N_t^{express}} \left(R_i^X - X_{t_i-}^{express} + D_i - \mathbb{E}[D_1]\right) + \mathbb{E}[D_1] N_t^{express}. \end{aligned} \tag{4.3.3}$$

The random variables $(R_i^X - X_{t_i-}^{express} + D_i - \mathbb{E}[D_i])_{i\geq 1}$ are i.i.d, bounded and with expectation 0 (see Appendix 4.6). We can then apply Hoeffding's inequality [Hoe63] and obtain the following lemma.

Lemma 4.3.10. *There exists $C_1 > 0$ such that for all $t \geq 0$ and $n \in \mathbb{N}\backslash\{0\}$, for all $k > 4n\mathcal{R}_{\tilde{\mu}}\mathbb{E}[D_1]^{-1}$, then*

$$\mathbb{P}\left(X_t^{express} < 4n\mathcal{R}_{\tilde{\mu}} \middle| N_t^{express} = k\right) \leq \exp\left(-\frac{\left(\mathbb{E}[D_1]k - 4n\mathcal{R}_{\tilde{\mu}}\right)^2}{C_1 k}\right).$$

Consequently, for all $C' > 4n\mathcal{R}_{\tilde{\mu}}\mathbb{E}[D_1]^{-1}$,

$$\mathbb{P}\left(X_t^{express} < 4n\mathcal{R}_{\tilde{\mu}} \middle| N_t^{express} > C'\right) \leq \exp\left(-\frac{\left(C'\mathbb{E}[D_1] - 4n\mathcal{R}_{\tilde{\mu}}\right)^2}{C_1 C'}\right).$$

Proof. First, notice that the second part of the lemma is a direct consequence of the first part, along with the variations of the function $x \to \exp(-(xc-d)^2 x^{-1})$, $c, d > 0$. As concerns the first part of the lemma, let $t \geq 0$, $n \in \mathbb{N}\backslash\{0\}$ and $k > 4n\mathcal{R}_{\tilde{\mu}}\mathbb{E}[D_1]^{-1}$. Using Eq. (4.3.3), we have

$$\begin{aligned} &\mathbb{P}\left(X_t^{express} < 4n\mathcal{R}_{\tilde{\mu}} \middle| N_t^{express} = k\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{k}(R_i^X - X_{t_i-}^{express} + D_i - \mathbb{E}[D_1]) < 4n\mathcal{R}_{\tilde{\mu}} - \mathbb{E}[D_1]k \middle| N_t^{express} = k\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{k}(-R_i^X + X_{t_i-}^{express} - D_i + \mathbb{E}[D_1]) > \mathbb{E}[D_1]k - 4n\mathcal{R}_{\tilde{\mu}} \middle| N_t^{express} = k\right). \end{aligned}$$

We conclude by using Hoeffding's inequality along with the fact that for all $i \geq 1$,

$$\begin{aligned} \left|R_i^X - X_{t_i-}^{express}\right| &< 4\mathcal{R}_{\tilde{\mu}} \\ \text{and} \quad |D_i - \mathbb{E}[D_1]| &\leq 2\mathcal{R}_{\tilde{\mu}}. \end{aligned}$$

□

In order to bound $\mathbb{E}[T_{0,n}^{express}]$, we also need to control the number of jumps made by the express chain over the time interval $[0, t]$.

Lemma 4.3.11. *There exists $C_\otimes > 0$ such that for all $t > 0$,*

$$\mathbb{P}\left(N_t^{express} \leq 0.1t\right) \leq \exp(-C_\otimes t).$$

Proof. The proof relies on the fact that the express chain jumps at rate 1. Hence, $N_t^{express}$ is Poisson distributed with parameter t.

Let $t > 0$. Using a Chernoff bound, we obtain

$$\begin{aligned}\mathbb{P}\left(N_t^{express} \leq 0.1t\right) &\leq \frac{(e\,t)^{0.1t}e^{-t}}{(0.1\,t)^{0.1t}} \\ &= \frac{\exp(0.1\,t)\exp(-t)}{\exp(0.1\,t\,\ln(0.1))} \\ &= \exp(0.1\,t - t - 0.1\ln(0.1)),\end{aligned}$$

and so taking $C_\otimes = 1 - 0.1(1 + \ln(10))$ allows us to conclude. □

Combining Lemmas 4.3.10 and 4.3.11, we obtain an upper bound for $\mathbb{E}[T_{0,n}^{express}]$.

Lemma 4.3.12. *There exists $C_2 > 0$ such that for all $n \in \mathbb{N}$,*

$$\mathbb{E}[T_{0,n}^{express}] \leq C_2 n.$$

Proof. Let $n \in \mathbb{N}$. Then,

$$\begin{aligned}\mathbb{E}[T_{0,n}^{express}] &= \int_0^\infty \mathbb{P}\left(T_{0,n}^{express} > t\right) dt \\ &= \int_0^t \mathbb{P}\left(X_t^{express} < 4n\mathcal{R}_{\tilde{\mu}}\right) dt \\ &\leq \frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]} + \int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^\infty \mathbb{P}\left(X_t^{express} < 4n\mathcal{R}_{\tilde{\mu}}\right) dt \\ &\leq \frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]} + \int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^\infty \mathbb{P}\left(N_t^{express} \leq \frac{t}{10}\right) dt \\ &\quad + \int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^\infty \mathbb{P}\left(\{X_t^{express} < 4n\mathcal{R}_{\tilde{\mu}}\} \cap \left\{N_t^{express} > \frac{t}{10}\right\}\right) \\ &\leq \frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]} + \int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^\infty \exp(-C_\otimes t) dt \\ &\quad + \int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^\infty \exp\left(-\frac{10}{C_1 t}\left(\frac{t\mathbb{E}[D_1]}{10} - 4n\mathcal{R}_{\tilde{\mu}}\right)^2\right) dt.\end{aligned} \tag{4.3.4}$$

Here we used Lemmas 4.3.10 and 4.3.11 to pass from the fourth to the fifth line.

Moreover,

$$\begin{aligned}
&\int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^{\infty} \exp\left(-\frac{10}{C_1 t}\left(\frac{t\mathbb{E}[D_1]}{10} - 4n\mathcal{R}_{\tilde{\mu}}\right)^2\right) dt \\
&= \exp\left(\frac{10}{C_1}\frac{2\mathbb{E}[D_1]}{10}4n\mathcal{R}_{\tilde{\mu}}\right) \\
&\quad \int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^{\infty} \exp\left(-\frac{10}{C_1 t}\frac{t^2\mathbb{E}[D_1]^2}{100}\right)\exp\left(-\frac{10}{C_1 t}16n^2\mathcal{R}_{\tilde{\mu}}^2\right) dt \\
&\leq \exp\left(\frac{8n}{C_1}\mathcal{R}_{\tilde{\mu}}\mathbb{E}[D_1]\right)\int_{\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}}^{\infty} \exp\left(-\frac{t}{10C_1}\mathbb{E}[D_1]^2\right) dt \\
&\leq \exp\left(\frac{8n}{C_1}\mathcal{R}_{\tilde{\mu}}\mathbb{E}[D_1]\right)\frac{10C_1}{\mathbb{E}[D_1]^2}\exp\left(-\frac{\mathbb{E}[D_1]^2}{10C_1}\frac{100n\mathcal{R}_{\tilde{\mu}}}{\mathbb{E}[D_1]}\right) \\
&\leq \frac{10C_1}{\mathbb{E}[D_1]^2}\exp\left(-\frac{2n}{C_1}\mathcal{R}_{\tilde{\mu}}\mathbb{E}[D_1]\right).
\end{aligned}$$

Since the first term in the r.h.s of (4.3.4) is proportional to n, and the second and third terms decrease exponentially fast in n, we can conclude. □

We can now conclude the proof of Lemma 4.3.7.

Proof. (Lemma 4.3.7) Let $n \in \mathbb{N}$. By Lemma 4.3.12,

$$\mathbb{E}[T_{0,n}^{express}] < +\infty,$$

and by Eq.(4.3.2),

$$\mathbb{E}[T_{0,n}] < +\infty.$$

□

We conclude this section with the proof of Proposition 4.3.1.

Proof. (Proposition 4.3.1) Since for all $n \in \mathbb{N}$, $T_{0,n} \geq 0$ and by Lemmas 4.3.12, 4.3.4, 4.3.5 and 4.3.6, the family $(T_{m,n})_{0\leq m\leq n}$ satisfies all the assumptions of Theorem 1.10 from [Lig85]. Therefore, there exists $\nu' \geq 0$ such that

$$\lim_{n\to+\infty}\frac{\mathbb{E}[T_{0,n}]}{n} = \frac{\mathbb{E}[\overleftarrow{\tau}_{4n\mathcal{R}_{\tilde{\mu}}}]}{n} = \nu'.$$

We conclude by standard upper and lower bounding arguments, using the fact that $x \to \mathbb{E}[\overleftarrow{\tau}_x]$ is a nondecreasing function (which is a consequence of Lemma 4.2.18). □

Remark 4.3.13. If we see $(X_t^{express})_{t\geq 0}$ as a cumulative process (in the sense of Chapter IV.3 from [Asm08]), it is possible to use Theorem 3.1 from this chapter and show that the limiting horizontal speed of advance of the express chain is equal to $\mathbb{E}[D_1]$, yielding an explicit lower bound on the speed of growth of the occupied area in the ∞-parent SLFV. In particular, if all reproduction ellipses have the same shape parameters (a, b, γ), which corresponds to the case investigated using numerical simulations, then the lower bound on the speed of growth is given by $\sqrt{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}$.

4.4 Upper bound on the speed of growth

We recall that $(E_t)_{t\geq 0}$ is the ∞-parent ancestral process with initial condition $\{(0,0)\}$ associated to $\tilde{\mu}^{\leftarrow}$, constructed using Π.

In this section, we complete the result shown in Section 4.3 by showing that the growth of the ∞-parent SLFV and of its dual counterpart are *at most* linear in time. This can be rewritten as a limiting property of $\mathbb{E}[\overleftarrow{\tau}_x]$ as follows.

Proposition 4.4.1.

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} > 0.$$

The proof can be found at the end of Section 4.4.3. Combining this result with Proposition 4.3.1 gives us the result of Theorem 4.2.11.

In order to show the result of Proposition 4.4.1, we first observe that it is sufficient to focus on the case in which all reproduction events are balls with fixed radius. Indeed, we have the following result.

Lemma 4.4.2. *Let $\Pi^{\mathcal{R}_{\tilde{\mu}}}$ be the Poisson point process defined over $\mathbb{R}_+ \times \mathbb{R}^d \times S_{\mathcal{R}_{\tilde{\mu}}}$ and with intensity*

$$\mu(S_{\mathcal{R}_{\tilde{\mu}}})dt \otimes dz \otimes \delta_{\mathcal{R}_{\tilde{\mu}}}(da) \otimes \delta_{\mathcal{R}_{\tilde{\mu}}}(db) \otimes \delta_0(\gamma).$$

Let $(E_t^{\mathcal{R}_{\tilde{\mu}}})_{t\geq 0}$ be the ∞-parent ancestral process with initial condition $\{(0,0)\}$ constructed using $\Pi^{\mathcal{R}_{\tilde{\mu}}}$, and for all $x>0$, let $\overleftarrow{\tau}^{\mathcal{R}_{\tilde{\mu}}}$ be the first time $(E_t^{\mathcal{R}_{\tilde{\mu}}})_{t\geq 0}$ reaches HP^x, defined as in (4.2.17). Then, for all $x>0$,

$$\mathbb{E}[\overleftarrow{\tau}_x] \geq \mathbb{E}\left[\overleftarrow{\tau}_x^{\mathcal{R}_{\tilde{\mu}}}\right].$$

Proof. The proof relies on the following coupling between $(E_t)_{t\geq 0}$ and $\left(E_t^{\mathcal{R}_{\tilde{\mu}}}\right)_{t\geq 0}$. Instead of being independent from Π, the Poisson point process $\Pi^{\mathcal{R}_{\tilde{\mu}}}$ is constructed using the points from Π, as follows: if $(t,z,a,b,\gamma) \in \Pi$, then $(t,z,\mathcal{R}_{\tilde{\mu}},\mathcal{R}_{\tilde{\mu}},0) \in \Pi^{\mathcal{R}_{\tilde{\mu}}}$. Since $\mathfrak{B}_{a,b,\gamma}(z) \subseteq \mathcal{B}_{\mathcal{R}_{\tilde{\mu}}}(z)$, this coupling ensures that

$$\forall t \geq 0, E_t \subseteq E_t^{\mathcal{R}_{\tilde{\mu}}} \text{ a.s.}$$

Therefore, for all $x>0$,

$$\min\{t\geq 0 : E_t \cap HP^x \neq \emptyset\} \geq \min\left\{t\geq 0 : E_t^{\mathcal{R}_{\tilde{\mu}}} \cap HP^x\right\} \text{ a.s.}$$

which allows us to conclude. □

Here we only provide the proof of Proposition 4.4.1 when reproduction events are balls with radius 1, but the proof can be generalized to balls of arbitrary fixed radius. We can then use Lemma 4.4.2 to obtain the corresponding result for ellipses with bounded parameters.

4.4.1 A first-passage percolation problem

We consider the graph $\mathcal{G}$ on the vertex set $\mathbb{Z}^2$, in which (i,j) and (i',j') are connected by an edge if, and only if

$$(i',j') \in \{(i+1,j),(i-1,j),(i,j+1),(i,j-1),(i-1,j-1),(i-1,j+1),(i+1,j-1),(i+1,j+1)\}.$$

To each edge e of $\mathcal{G}$, we associate a random variable

$$\mathcal{E}_e \sim \mathcal{E}xp(16\times\pi^{-1}) = \mathcal{E}xp(16\times V_1^{-1}),$$

where we recall that V_1 is the volume of a ball with radius 1. $\mathcal{E}_e$ corresponds to the time needed to pass through the corresponding edge. Following standard terminology in first-passage percolation, we call it the *passage time* of the edge. The choice of the rate of the exponential distribution ensures we can later compare the growth of the ∞-parent ancestral process to the first-passage percolation problem we now introduce.

If Γ is a (potentially infinite) path formed by the edges $e_1, ..., e_n, ...$, then the passage time of the path Γ is defined as

$$\mathcal{E}_\Gamma = \sum_{e \in \Gamma} \mathcal{E}_e.$$

If $z_1, z_2 \in \mathbb{Z}^2$, we define the first-passage time $\mathcal{E}_{z_1,z_2}$ from z_1 to z_2 (or equivalently, from z_2 to z_1 since $\mathcal{G}$ is not oriented) as the minimum over the passage times of all the (finite) paths going from z_1 to z_2. We then define

$$\overleftarrow{\tau}_n^{fpp} := \min\left\{t \geq 0 : \exists m \in \mathbb{Z}, \mathcal{E}_{(0,0),(n,m)} \leq t\right\}.$$

In other words, $\overleftarrow{\tau}_n^{fpp}$ is the time needed to reach a point at horizontal separation n from the origin, starting from the origin.

The interest of this first-passage percolation problem lies in the fact that since the passage time of any given edge is almost surely stricty positive, generalizing Theorem 6.7 from [SW78] to our lattice yields the following result.

Lemma 4.4.3.

$$\lim_{n \to +\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_n^{fpp}]}{n} > 0.$$

In order to use this lemma and show Proposition 4.4.1, we need to be able to compare $\overleftarrow{\tau}_x$ and $\overleftarrow{\tau}_n^{fpp}$. The main obstacle to this comparison lies in the fact that the ∞-parent ancestral process is continuous in space, while the first-passage percolation problem is defined on a graph. Therefore, we now introduce a way to "discretize" the ∞-parent ancestral process.

4.4.2 Discretization of the ∞-parent ancestral process

In order to discretize the ∞-parent ancestral process, we first place a grid on $\mathbb{R}^2$, and associate a cell to each site of the lattice. Let

$$\mathcal{V} := \left\{(4i, 4j) : (i, j) \in \mathbb{Z}^2\right\}$$

be the underlying grid, and for all $(i, j) \in \mathbb{Z}^2$, let $\mathcal{C}_{i,j}$ be the square with center $(4i, 4j)$ and side length 2. That is,

$$\mathcal{C}_{i,j} := \left\{(x, y) \in \mathbb{R}^2 : |x - 4i| \leq 1 \text{ and } |y - 4j| \leq 1\right\}.$$

Each $\mathcal{C}_{i,j}$, $i, j \in \mathbb{Z}$ corresponds to the cell associated to the site $(4i, 4j)$ of the grid $\mathcal{V}$. See Figure 4.7 for an illustration.

This construction satisfies the two following key properties.

1. For all $z = (x, y) \in \mathbb{R}^2$, unless $x = 4i + 2$, $i \in \mathbb{Z}$ or $y = 4j + 2$, $j \in \mathbb{Z}$, the ball $\mathcal{B}_1(z)$ intersects at most one cell.

   In other words, each reproduction events intersects almost surely at most one cell. Moreover, if $\mathcal{B}_1(z)$ does not intersect any cell, then it means that z fell into one of the areas in white on Figure 4.7.

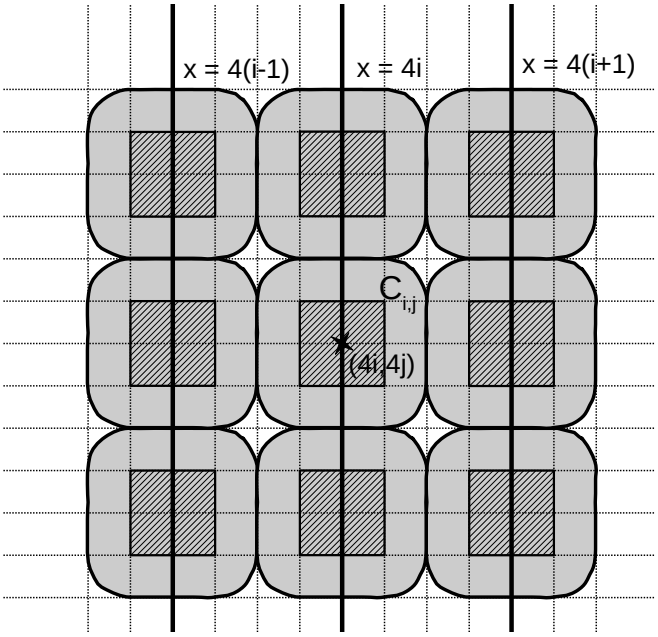


Figure 4.7: Grid used to discretize the ∞-parent ancestral process when reproduction events are balls. Each hatched square corresponds to a cell. The grey area around the site $(4i, 4j)$, indicated by a black cross, corresponds to the area in which centers of reproduction events intersecting the cell $\mathcal{C}_{i,j}$ can fall. The white areas contain all the centers of reproduction events which do not intersect any cell.

2. If we refer to a sequence of reproduction events occurring in chronological order and such that each reproduction event intersects the previous one as a *path of reproduction events*, then any path of reproduction events for which none of the corresponding balls intersect a cell is almost surely confined in one of the white areas on Figure 4.7, in the sense that all the reproduction event centers fall into the same white area almost surely.

The discretized version of the ∞-parent ancestral process, denoted $(D_t)_{t\geq 0}$, is then defined as follows. For all $t \geq 0$, let

$$D_t := \left\{(i,j) \in \mathbb{Z}^2 : \mathcal{C}_{i,j} \cap E_t \neq \emptyset\right\}$$

be the set of all cells which intersect the ∞-parent ancestral process with parameters $(a, b, \gamma) = (1, 1, 0)$ and initial condition $(0, 0)$ at time t. Moreover, we associate to $(D_t)_{t\geq 0}$ the random variables $(\overleftarrow{\tau}_n^{discr})_{n\geq 0}$ such that for all $n \geq 0$,

$$\overleftarrow{\tau}_n^{discr} := \min\left\{t \geq 0 : \exists m \in \mathbb{Z}, (n,m) \in D_t\right\}.$$

Due to the structure of the grid and the size of the cells, we have the following result.

Lemma 4.4.4. *For all* $n \in \mathbb{N}\backslash\{0\}$*,*

$$\overleftarrow{\tau}_n^{discr} \leq \overleftarrow{\tau}_{4n} \quad \textit{a.s.}$$

Proof. Let $n \in \mathbb{N}\backslash\{0\}$. Let $z \in \mathbb{R}^2$ be the center of the reproduction event which occurs at time $\overleftarrow{\tau}_{4n}$ and makes the ∞-parent ancestral process reach for the first time a point at horizontal separation $4n$ from the origin. Then, there exists almost surely $j \in \mathbb{Z}$ such that

$$\mathcal{B}_1(z) \cap \mathcal{C}_{n,j} \neq \emptyset,$$

and hence $\overleftarrow{\tau}_n^{discr} \leq \overleftarrow{\tau}_{4n}$. □

In all that follows, for all $(i,j),(i',j') \in \mathbb{Z}^2$, we say that

- the cell $\mathcal{C}_{i,j}$ is *active* at time t if $(i,j) \in D_t$;
- the cell $\mathcal{C}_{i,j}$ is *activated* at time t if $(i,j) \in D_t$ and $(i,j) \notin D_{t-}$;
- the cell $\mathcal{C}_{i',j'}$ *activates* $\mathcal{C}_{i,j}$ at time t if there exists $s < t$ such that $\mathcal{C}_{i',j'}$ is active at time s, and if there exists a path of reproduction events starting from $\mathcal{C}_{i',j'}$ at time s, initially overlapping an area of $\mathcal{C}_{i',j'}$ containing type 1 individuals, and reaching $\mathcal{C}_{i,j}$ for the first time at time t while not intersecting any other cell on the time interval $[s,t]$.

Notice that under this terminology, a cell can activate another one which is already active. Moreover, the only cells that the cell $\mathcal{C}_{i',j'}$ can activate are (almost surely) its nearest neighbours in the graph $\mathcal{G}$, that is, the cells $\mathcal{C}_{i,j}$ such that

$$\sqrt{(i-i')^2+(j-j')^2} \leq \sqrt{2}.$$

4.4.3 Comparison to the first-passage percolation problem

We now compare the growth of the discretized ∞-parent ancestral process $(D_t)_{t\geq 0}$ to the one of the first-passage percolation problem introduced earlier. We recall that

$$\overleftarrow{\tau}_n^{fpp} := \min\left\{t \geq 0 : \exists m \in \mathbb{Z}, \mathcal{E}_{(0,0),(n,m)} \leq t\right\}$$

is the time needed by the process associated to the first-passage percolation problem to reach a point at horizontal separation n from the origin, starting from the origin.

Proposition 4.4.5. *$\overleftarrow{\tau}_n^{discr}$ is stochastically bounded from below by $\overleftarrow{\tau}_n^{fpp}$, that is, for all $t \geq 0$,*

$$\mathbb{P}\left(\overleftarrow{\tau}_n^{discr} \geq t\right) \geq \mathbb{P}\left(\overleftarrow{\tau}_n^{fpp} \geq t\right).$$

Proof. We recall that for all $e \in \mathcal{G}$, $\mathcal{E}_e \sim \mathcal{E}xp(16\pi^{-1})$ is the passage time of edge e. In order to show Proposition 4.4.5, we show that cells are activated faster in the first-passage percolation problem than in the discretized ∞-parent ancestral process, and conclude by induction on the number of cells reached.

First, we observe that for both processes:

- cells are activated one after another a.s.,
- a cell cannot be activated if all its neighbors are inactive.

Therefore, we can focus on the time needed for a cell to become active once (a.s. exactly) one of its neighbors become active. Regarding the first-passage percolation process, this time is bounded from above by $\mathcal{E}_e$, where e is the edge connecting the active neighboring cell to the focal cell. Regarding the discretized ∞-parent ancestral process, this time is bounded from below by the time needed for the cell to be intersected by a reproduction event (which is a prerequisite for the cell to become active). Such reproduction events occur at a rate bounded from above by $16\pi^{-1}$. Therefore, the time needed for the cell to become active in the discretized ∞-parent ancestral process is stochastically bounded from below by $\mathcal{E}_e$. □

Remark 4.4.6. Due to correlations between activations by neighbouring cells, it woud be more difficult to construct a coupling with the first-passage percolation problem. However, the stochastic comparison we obtained is sufficient.

We can now show Proposition 4.4.1.

Proof. (Proposition 4.4.1) By Proposition 4.4.5 and Lemma 4.4.4, for all $n \in \mathbb{N}\backslash\{0\}$,

$$\mathbb{E}[\overleftarrow{\tau}_{4n}] \geq \mathbb{E}[\overleftarrow{\tau}_n^{discr}] \geq \mathbb{E}[\overleftarrow{\tau}_n^{fpp}].$$

By Lemma 4.4.3, we know that

$$\lim_{n\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_n^{fpp}]}{4n} > 0.$$

Since we know that $\lim_{n\to+\infty} \mathbb{E}[\overleftarrow{\tau}_n] \times n^{-1}$ exists by Proposition 4.3.1, we obtain that

$$\lim_{n\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_{4n}]}{4n} > 0.$$

We conclude by using the fact that $x \to \overleftarrow{\tau}_x$ is nondecreasing by Lemma 4.2.18. □

We conclude this section with the proof of Theorem 4.2.11.

Proof. (Theorem 4.2.11) By Proposition 4.3.1, we know that there exists $\nu \geq 0$ such that

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} = \nu.$$

Moreover, by Proposition 4.4.1, we know that

$$\lim_{x\to+\infty} \frac{\mathbb{E}[\overleftarrow{\tau}_x]}{x} > 0.$$

Therefore, $\nu > 0$ and we conclude using Proposition 4.2.20. □

4.5 The two column growth process

The numerical simulations performed in Section 4.2.4 show that the speed of growth of the occupied region in the ∞-parent SLFV is higher than the one conjectured using the express chain. Moreover, the results suggest that the growth of the process is driven by relatively unfrequent "spikes" than then thicken and grow sideways. In order to investigate how these spikes can significantly increase the growth speed, we introduce the following toy model. We consider two adjacent piles of cubes, thereafter referred to as the *left pile* and the *right pile*. Each cube has height 1. A cube is added on top of the left (resp. right) pile at rate 1, independently from the other pile. Moreover, the piles can also grow sideways: if there is a cube at height h in the left (resp. right) pile, and no cube at such height in the other pile, then a cube is added to the right (resp. left) pile at height h at rate 1. In particular, if the left pile is one cube higher than the right (resp. left) pile, then the total rate at which the height of the right (resp. left) pile increases of 1 is equal to 2, as the growth can be due to a new cube falling on top of the right (resp. left) pile as well as sideways growth of the left (resp. right) pile. See Figure 4.8 for an illustration of the dynamics of the process.

Remark 4.5.1. The term "pile" can be a bit misleading, since it is possible to have holes in it, due to the other pile growing sideways. See Figure 4.8 for an illustration.

The interest of the process lies in the following observation. If l_t (resp. r_t) represents the maximal height reached by the left (resp. right) pile, then the cubes at height $h < \min(l_t, r_t)$ no longer contribute to the growth of the process. Therefore, we can "reset" the process whenever $l_t = r_t$.

We now define this process, called the *two columns growth process*, rigorously.

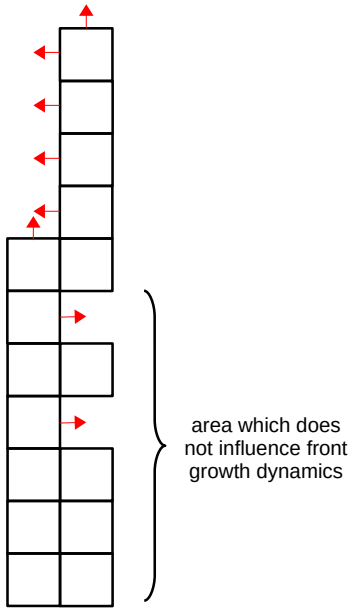


Figure 4.8: Growth dynamics of the two columns growth process. The red arrows indicate possible growth events occurring at rate 1.

4.5.1 The two column growth process: Definition and properties

The state space over which the process is defined is the set $\widetilde{\mathcal{S}}$ defined by

$$\widetilde{\mathcal{S}} := \{(i,j) \in \mathbb{N} \times \mathbb{N} : i \leq j\}.$$

If $(i,j) \in \widetilde{\mathcal{S}}$, then i represents the height reached by the *lowest* pile, and j the height reached by the *highest* pile.

Definition 4.5.2. *Let $(i,j) \in \widetilde{\mathcal{S}}$. The two columns growth process $(G_t)_{t\geq 0} = (m_t, M_t)_{t\geq 0}$ with initial condition (i,j), thereafter referred to as the 2-CGP, is the continuous-time Markov chain on the state space $\widetilde{\mathcal{S}}$ whose transition rates are as follows. For all $(i,j) \in \widetilde{\mathcal{S}}$,*

1. *If $i = j$, then $(i,j) \to (i,i+1)$ at rate 2, and no other transitions are possible.*
2. *If $j \geq i+1$, then*

$$(i,j) \to \begin{cases} (i,j+1) & \text{at rate } 1, \\ (i+1,j) & \text{at rate } 2, \\ (i+k,j), k \in [\![2, j-i]\!] & \text{at rate } 1 \text{ if } j > i+1, \end{cases}$$

and no other transitions are possible.

These transition rates exactly encode the dynamics described earlier. Indeed, if $i = j$, then both piles have the same height, and cubes fall onto each one of the piles at rate 1, yielding a total transition rate of 2 as we do not record which pile is the highest. If $i \neq j$, then each pile can grow upwards at rate 1, but the highest pile can also grow sideways. Depending on where the highest pile grows sideways, this results in a more or less sharp increase in the height of the lowest pile.

Earlier, we saw that it was possible to "reset" the process whenever the two piles had the same height. In order to do so, let $\mathcal{S}_{\square\square}$ be the set

$$\mathcal{S}_{\square\square} := \{(i,i) : i \in \mathbb{N}\}$$

of all configurations such that both piles have the same height, and let $T_\square$ be the time of first return of $(G_t)_{t\geq 0}$ to a state belonging to $\mathcal{S}_{\square\square}$. Our first result is an upper bound on $\mathbb{E}_{(0,0)}[T_\square]$.

Lemma 4.5.3.

$$\mathbb{E}_{(0,0)}[T_\square] \leq 3/2.$$

Proof. In order to do so, we first notice that $T_\square$ is equal to the sum of the time needed to exit $(0,0)$, which follows an exponential distribution with parameter 2, and of the time needed to reach a state in $\mathcal{S}_{\square\square}$ starting from $(0,1)$.

The first step yields an expected contribution of $1/2$ to $\mathbb{E}_{(0,0)}[T_\square]$. Once the process is in the state $(0,1)$, assuming without loss of generality that the cube fell on the right pile (which is now the highest pile), we assign an exponential clock with parameter 1 to the highest cube of the right pile, which rings whenever the cube attempts to grow sideways. Whenever the highest cube changes, the exponential clock is assigned to the new highest cube. When the clock rings for the first time, we distinguish two cases:

- Either there is no cube at the same height in the left pile, and the cube can grow sideways. Then, we are back to a state in $\mathcal{S}_{\square\square}$, though perhaps not for the first time.
- Either there is already a cube at the same height in the left pile. Then, we know we already went back to a state in $\mathcal{S}_{\square\square}$.

Therefore,

$$\mathbb{E}_{(0,0)}[T_\square] \leq 1/2 + 1,$$

allowing us to conclude. □

We can use this lemma to obtain an upper bound on $\mathbb{E}_{(0,0)}\left[M_{T_\square}\right]$ and $\mathbb{E}_{(0,0)}\left[m_{T_\square}\right]$.

Lemma 4.5.4.

$$\mathbb{E}_{(0,0)}\left[M_{T_\square}\right] \leq 2 \quad \textit{and} \quad \mathbb{E}_{(0,0)}\left[m_{T_\square}\right] \leq 2.$$

Proof. First, as $G_{T_\square} \in \mathcal{S}_{\square\square}$, we have $M_{T_\square} = m_{T_\square}$, and it is sufficient to provide an upper bound on $\mathbb{E}_{(0,0)}\left[M_{T_\square}\right]$. Then, we set

$$\widetilde{T} := \min\{t \geq 0 : M_t \neq m_t\},$$

which corresponds to the time needed for the process to leave the state $(0,0)$. For all $t \in [\widetilde{T}, T_\square)$, $M_t > m_t$, and the lowest pile does not contribute to the growth of the highest pile. Therefore, over the time interval $[\widetilde{T}, T_\square)$, the only way the highest pile can grow is by new cubes falling on top of it, whose total number is given by $M_{T_\square} - M_{\widetilde{T}} = M_{T_\square} - 1$. Moreover, conditional on $T_\square - \widetilde{T}$, this number follows a Poisson distribution with parameter $T_\square - \widetilde{T}$. Therefore,

$$\begin{aligned}
\mathbb{E}_{(0,0)}\left[M_{T_\square}\right] &= \mathbb{E}_{(0,0)}\left[M_{\widetilde{T}}\right] + \mathbb{E}_{(0,0)}\left[M_{T_\square} - M_{\widetilde{T}}\right] \\
&= 1 + \mathbb{E}_{(0,0)}\left[\mathbb{E}_{(0,0)}\left[M_{T_\square} - M_{\widetilde{T}} \middle| T_\square - \widetilde{T}\right]\right] \\
&= 1 + \mathbb{E}_{(0,0)}\left[T_\square - \widetilde{T}\right] \\
&= 1 - 1/2 + \mathbb{E}_{(0,0)}[T_\square] \\
&\leq 1/2 + 3/2.
\end{aligned}$$

by Lemma 4.5.3, which allows us to conclude. □

In order to study the speed of growth of the process, we see the 2-CGP as a cumulative process, in the sense of Chapter VI from [Asm08]. We obtain the following theoretical result on the speed of growth of the process.

Theorem 4.5.5. *In the notation introduced earlier,*

$$\lim_{t\to+\infty}\frac{M_t}{t}=\lim_{t\to+\infty}\frac{m_t}{t}=\frac{\mathbb{E}_{(0,0)}\left[M_{T_\square}\right]}{\mathbb{E}_{(0,0)}\left[T_\square\right]}\ \textit{a.s.}$$

Proof. Let $T_\square^0=0$, and for all $n\geq 1$, let $T_\square^n$ be the time of n-th return of $(G_t)_{t\geq 0}$ to a state in $\mathcal{S}_{\square\square}$. For all $t>0$, let $N(t)=\max\{n\in\mathbb{N}:T_\square^n\leq t\}$. We can rewrite m_t and M_t as

$$m_t=\sum_{n=1}^{N(t)}\left(M_{T_\square^n}-M_{T_\square^n-1}\right)+m_t-M_{T_\square^{N(t)}}$$
$$\text{and } M_t=\sum_{n=1}^{N(t)}\left(M_{T_\square^n}-M_{T_\square^n-1}\right)+M_t-M_{T_\square^{N(t)}}.$$

Here we used the fact that for all $n\in\mathbb{N}$, $M_{T_\square^n}=m_{T_\square^n}$.

We observe that the random variables $\left(M_{T_\square^n}-M_{T_\square^{n-1}}\right)_{n\geq 1}$ are i.i.d. The same is true for the random variables $(T_\square^n-T_\square^{n-1})_{n\geq 1}$. If we can show that

$$\mathbb{E}_{(0,0)}\left[\max_{0\leq t\leq T_\square}M_t\right]<+\infty \tag{4.5.1}$$
$$\text{and } \mathbb{E}_{(0,0)}\left[\max_{0\leq t\leq T_\square}m_t\right]<+\infty,. \tag{4.5.2}$$

then we can use Theorem 3.1 from Chapter VI of [Asm08] and conclude. As $t\to M_t$ and $t\to m_t$ are non-decreasing, this amounts to showing that $\mathbb{E}_{(0,0)}\left[M_{T_\square}\right]<+\infty$ and $\mathbb{E}_{(0,0)}\left[m_{T_\square}\right]<+\infty$, which is a direct consequence of Lemma 4.5.4. □

In order to use this result and obtain the limiting speed of growth of the 2-CGP, we need to be able to compute $\mathbb{E}_{(0,0)}[M_{T_\square}]$ and $\mathbb{E}_{(0,0)}[T_\square]$. As a first step, we introduce the martingale problem satisfied by the 2-CGP, which we will use to obtain a relation between $\mathbb{E}_{(0,0)}[M_{T_\square}]$ and $\mathbb{E}_{(0,0)}[T_\square]$.

Let $C_b(\widetilde{\mathcal{S}})$ be the space of bounded functions $f:\widetilde{\mathcal{S}}\to\mathbb{R}$. The generator $\mathcal{G}$ of the 2-CGP acting on functions $f\in C_b(\widetilde{\mathcal{S}})$ is defined as follows. For all $f\in C_b(\widetilde{\mathcal{S}})$ and for all $(i,j)\in\widetilde{\mathcal{S}}$),

$$\begin{aligned}\mathcal{G}f(i,j)=&\mathbb{1}_{\{i=j\}}\times 2\left(f(i,i+1)-f(i,j)\right)\\&+\mathbb{1}_{\{i+1=j\}}\times\left[2(f(i+1,j)-f(i,j))+f(i,j+1)-f(i,j)\right]\\&+\mathbb{1}_{\{j\geq i+2\}}\times\left[2(f(i+1,j)-f(i,j))+f(i,j+1)-f(i,j)+\sum_{k=2}^{j-i}(f(i+k,j)-f(i,j))\right].\end{aligned}$$

The 2-CGP is then a solution to the following martingale problem.

Lemma 4.5.6. *Let $(i,j)\in\widetilde{\mathcal{S}}$, and let $(G_t)_{t\geq 0}=(m_t,M_t)_{t\geq 0}$ be the 2-CGP with initial condition (i,j). Then, for all $f\in C_b(\widetilde{\mathcal{S}})$,*

$$\left(f(G_t)-f((i,j))-\int_0^t\mathcal{G}f(G_s)ds\right)_{t\geq 0}$$

is a martingale.

We use the martingale problem with functions of the form $f_d : (i,j) \to j\mathbb{1}_{\{j<d\}}$, $d \in \mathbb{N}\backslash\{0,1\}$. Indeed, for all $d \in \mathbb{N}\backslash\{0,1\}$ and $s \geq 0$, if $G_s = (m_s, M_s)$ is such that $M_s - m_s \geq 1$, then

$$\mathcal{G}f_d(G_s) = \mathbb{1}_{\{M_s<d-1\}} - \mathbb{1}_{\{M_s=d-1\}}(d-1) \tag{4.5.3}$$

and if $m_s = M_s$, then

$$\mathcal{G}f_d(G_s) = 2\mathbb{1}_{\{M_s<d-1\}} - 2(d-1)\mathbb{1}_{\{M_s=d-1\}}. \tag{4.5.4}$$

We obtain the following result.

Lemma 4.5.7. *Under the notation of Theorem 4.5.5,*

$$\mathbb{E}_{(0,0)}[M_{T_\square}] = \frac{1}{2} + \mathbb{E}_{(0,0)}[T_\square].$$

Proof. For all $d \in \mathbb{N}\backslash\{0,1\}$, let $T_d := \inf\{t \geq 0 : M_t \geq d\}$. We use the martingale problem stated in Lemma 4.5.6 with the function $f_d : (i,j) \to j\mathbb{1}_{\{j<d\}}$, $d \in \mathbb{N}\backslash\{0,1\}$, and the stopping time $T_\square \wedge T_d$. We obtain

$$\begin{aligned}
\mathbb{E}_{(0,0)}\left[M_{T_\square\wedge T_d}\right] &= \mathbb{E}_{(0,0)}\left[M_{T_\square\wedge T_d}\mathbb{1}_{\{M_{T_\square\wedge T_d}<d+2\}}\right] \\
&= 0 + \mathbb{E}_{(0,0)}\left[\int_0^{T_d\wedge T_d} \mathcal{G}f_{d+2}(G_s)ds\right] \\
&= \mathbb{E}_{(0,0)}\left[\int_0^{T_1} \mathcal{G}f_{d+2}(G_s)ds\right] + \mathbb{E}_{(0,0)}\left[\int_{T_1}^{T_\square\wedge T_d} \mathcal{G}f_{d+2}(G_s)ds\right] \\
&= \mathbb{E}_{(0,0)}\left[2T_1\right] + \mathbb{E}_{(0,0)}\left[T_\square \wedge T_d - T_1\right] \\
&= \mathbb{E}_{(0,0)}\left[T_1\right] + \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_\square\leq T_d\}}\right] + \mathbb{E}_{(0,0)}\left[T_d\mathbb{1}_{\{T_d<T_\square\}}\right] \\
&= \frac{1}{2} + \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_\square\leq T_d\}}\right] + \mathbb{E}_{(0,0)}\left[T_d\mathbb{1}_{\{T_d<T_\square\}}\right].
\end{aligned}$$

Here we used (4.5.3) and (4.5.4) to pass from the second to the third line. Moreover, we also have

$$\mathbb{E}_{(0,0)}\left[M_{T_\square\wedge T_d}\right] = \mathbb{E}_{(0,0)}\left[M_{T_\square}\mathbb{1}_{\{T_\square\leq T_d\}}\right] + \mathbb{E}_{(0,0)}\left[d\mathbb{1}_{\{T_\square>T_d\}}\right].$$

Therefore, if we show that

$$\lim_{d\to+\infty} \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_\square\leq T_d\}}\right] = \mathbb{E}_{(0,0)}\left[T_\square\right], \tag{4.5.5}$$

$$\lim_{d\to+\infty} \mathbb{E}_{(0,0)}\left[T_d\mathbb{1}_{\{T_d<T_\square\}}\right] = 0, \tag{4.5.6}$$

$$\lim_{d\to+\infty} \mathbb{E}_{(0,0)}\left[M_{T_\square}\mathbb{1}_{\{T_\square\leq T_d\}}\right] = \mathbb{E}_{(0,0)}\left[M_{T_\square}\right], \tag{4.5.7}$$

$$\text{and} \quad \lim_{d\to+\infty} \mathbb{E}_{(0,0)}\left[M_{T_d}\mathbb{1}_{\{T_d<T_\square\}}\right] = 0, \tag{4.5.8}$$

then we will be able to conclude.

In order to do so, we recall that in order to come back to the set $\mathcal{S}_{\square\square}$, the process first needs to leave the state $(0,0)$, which occurs at time $T_1 \sim \mathcal{E}xp(2)$. Without loss of generality, we assume that the first cube falls on the right pile. As in the proof of Lemma 4.5.3, we assign an exponential clock $T_\to \sim \mathcal{E}xp(1)$ to the highest cube of the right pile, and move this exponential clock to the new highest cube in this pile whenever it grows. Then, reasoning as in the proof of Lemma 4.5.3, $T_\square \leq T_\to + T_1$. Moreover, $M_{T_\square} \leq d$ if, and only if at most $d-1$ cubes fall on the right pile during the time interval

$[T_1, T_\square]$. Therefore, if $\mathcal{P}(\lambda)$, $\lambda > 0$ stands for a Poisson random variable with parameter λ,

$$\begin{aligned}\mathbb{P}_{(0,0)}(T_\square > T_d) &\leq \mathbb{P}(\mathcal{P}(T_\rightarrow) \geq d-1)\\ &\leq \int_0^\infty e^{-t}\mathbb{P}(\mathcal{P}(t) \geq d-1)dt\\ &\leq \int_0^\infty e^{-t}\frac{\mathbb{E}[\mathcal{P}(t)]}{d-1}dt\\ &\leq \frac{1}{d-1}\int_0^\infty te^{-t}dt\\ &\xrightarrow[d\to+\infty]{} 0.\end{aligned}$$

Therefore, by the dominated convergence theorem and Lemmas 4.5.3, 4.5.4,

$$\lim_{d\to+\infty}\mathbb{E}_{(0,0)}\left[T_\square \mathbb{1}_{\{T_d<T_\square\}}\right] = 0$$
$$\text{and}\quad \lim_{d\to+\infty}\mathbb{E}_{(0,0)}\left[M_{T_\square}\mathbb{1}_{\{T_d<T_\square\}}\right] = 0,$$

from which we deduce (4.5.5) and (4.5.7). Moreover,

$$\mathbb{E}_{(0,0)}\left[T_d\mathbb{1}_{\{T_d<T_\square\}}\right] \leq \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_d<T_\square\}}\right],$$

giving (4.5.6), and as $t \to M_t$ is non-decreasing,

$$\mathbb{E}_{(0,0)}\left[M_{T_d}\mathbb{1}_{\{T_d<T_\square\}}\right] \leq \mathbb{E}_{(0,0)}\left[M_{T_\square}\mathbb{1}_{\{T_d<T_\square\}}\right],$$

allowing us to conclude, again by a dominated convergence argument. □

Using this result along with Theorem 4.5.5, we obtain a new expression for the speed of growth of the process, along with explicit lower and upper bounds.

Proposition 4.5.8.

$$\lim_{t\to+\infty}\frac{M_t}{t} = 1 + \frac{1}{2\mathbb{E}_{(0,0)}[T_\square]} \in [4/3, 2] \textit{ a.s.}$$

Proof. By Theorem 4.5.5,

$$\lim_{t\to+\infty}\frac{M_t}{t} = \frac{\mathbb{E}_{(0,0)}\left[M_{T_\square}\right]}{\mathbb{E}_{(0,0)}\left[T_\square\right]} \text{ a.s.}$$

Moreover, by Lemma 4.5.7,

$$\begin{aligned}\frac{\mathbb{E}_{(0,0)}\left[M_{T_\square}\right]}{\mathbb{E}_{(0,0)}\left[T_\square\right]} &= \frac{1/2 + \mathbb{E}_{(0,0)}\left[T_\square\right]}{\mathbb{E}_{(0,0)}\left[T_\square\right]}\\ &= 1 + \frac{1}{2\mathbb{E}_{(0,0)}\left[T_\square\right]}.\end{aligned}$$

By Lemma 4.5.3, $\mathbb{E}_{(0,0)}\left[T_\square\right] \leq 3/2$ so

$$1 + \frac{1}{2\mathbb{E}_{(0,0)}\left[T_\square\right]} \geq 1 + 1/3 = 4/3.$$

Moreover, as $T_\square \geq T_1$ and as $T_1 \sim \mathcal{E}xp(2)$, $\mathbb{E}_{(0,0)}\left[T_\square\right] \geq 1/2$. Therefore,

$$1 + \frac{1}{2\mathbb{E}_{(0,0)}\left[T_\square\right]} \leq 1 + 1 = 2,$$

allowing us to conclude. □

Remark 4.5.9. Without the interaction with the other pile, the speed of growth of an isolated pile of cubes would be of 1. Therefore, this first result means that the interaction between the two piles increases the speed of growth by a factor of at least 1.33, and at most 2. Considering more than two interacting piles would increase even more the speed of growth, and yield a factor closer to the one obtained for the SLFV (whose value of 2.58 was obtained through numerical simulations in Section 4.2.4).

4.5.2 The discretized two columns growth process

The main obstacle to the study of the speed of growth of the 2-CGP lies in the fact that the process does not jump at a constant rate: the bigger the height difference between the two columns, the faster the process jumps. In order to circumvent this problem, we now introduce a discretized version of the two columns growth process. Then, we explain how to couple it to the 2-CGP, and use its invariant distribution to obtain an approximation of $\mathbb{E}_{(0,0)}[T_{\square}]$ and of the speed of growth. In all that follows, let $(G_t)_{t\geq 0} = (m_t, M_t)_{t\geq 0}$ be the 2-CGP with initial condition $(0,0)$. We recall that $T_{\square}$ is the time of first return of $(G_t)_{t\geq 0}$ to a state in $\mathcal{S}_{\square\square}$, and for all $d \in \mathbb{N}\backslash\{0,1\}$, $T_d = \inf\{t \geq 0 : M_t \geq d\}$.

Moreover, let $N \in \mathbb{N}\backslash\{0,1\}$ and $\epsilon > 0$. In order to construct the discretized 2-CGP, we make the following observation. If $t \geq 0$, then the probability that the process $(G_t)_{t\geq 0}$ jumps at least once in the time interval $[t, t+\epsilon)$ is equal to

$$1 - \exp\left(-\epsilon(M_t - m_t + 2)\right) \approx \epsilon(M_t - m_t + 2),$$

and the probability that it jumps at least twice is bounded from above by

$$\left(1 - \exp\left(-\epsilon(M_t - m_t + 3)\right)\right)^2 \approx \epsilon^2(M_t - m_t + 3)^2.$$

Therefore, if ϵ is small enough and if we are able to control $M_t - m_t$, then we can consider that at most one growth event occurs on a time interval of length ϵ. We use this idea to construct the discretized 2-CGP. In order to ease the notation, we only describe the dynamics of the height difference between the two piles, and define the discretized 2-CGP on the state space $\mathbb{N}$. Note that it is possible to recover the complete process from the evolution of the height difference. Moreover, we will often abuse notation and say that the discretized 2-CGP starts from the state $(0,0)$, or that it comes back to a state in $\mathcal{S}_{\square\square}$ when it comes back to the state 0.

Definition 4.5.10. *For all $\epsilon > 0$ and $N \in \mathbb{N}\backslash\{0,1\}$, the discretized 2-CGP $(\hat{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ with time step ϵ and maximal height difference N is the $[\![0,N]\!]$-valued discrete-time Markov chain with initial condition $\hat{G}_0^{(N,\epsilon)} = 0$ and whose transition probabilities $(\hat{p}_{i,j}^{(N,\epsilon)})_{(i,j)\in[\![0,N]\!]^2}$ are defined as follows.*

1. *If $i = 0$, $\hat{p}_{0,0}^{(N,\epsilon)} = \exp(-2\epsilon)$, $\hat{p}_{0,1}^{(N,\epsilon)} = 1 - \exp(-2\epsilon)$, and for all $j \in [\![2,N]\!]$, $\hat{p}_{0,j}^{(N,\epsilon)} = 0$.*
2. *For all $i \in [\![1, N-1]\!]$,*

$$\hat{p}_{i,j}^{(N,\epsilon)} := \begin{cases} 0 & \text{if } j > i+1, \\ \frac{2}{i+2}(1 - \exp(-(i+2)\epsilon)) & \text{if } j = i-1, \\ \exp(-(i+2)\epsilon) & \text{if } j = i, \\ \frac{1}{i+2}(1 - \exp(-(i+2)\epsilon)) & \text{if } j = i+1 \text{ or (if } i \neq 1) \ 0 \leq j \leq i-2, \end{cases}$$

3. *If $i = N$,*

$$\hat{p}_{N,j}^{(N,\epsilon)} := \begin{cases} 0 & \text{if } j > N+1, \\ \frac{2}{N+2}(1 - \exp(-(N+2)\epsilon)) & \text{if } j = N-1, \\ 1 - \frac{N+1}{N+2}(1 - \exp(-(N+2)\epsilon)) & \text{if } j = N, \\ \frac{1}{N+2}(1 - \exp(-(N+2)\epsilon)) & \text{if } 0 \leq j \leq N-2, \end{cases}$$

This process has dynamics similar to the ones of the 2-CGP, except when the height difference between the two piles is equal to N: the growth of the highest pile in then blocked until the lower piles grows. This ensures we can compute the invariant distribution of the process.

Before studying the properties of the discretized 2-CGP, we explain how to couple it to the (continuous-time) 2-CGP. In order to do so, for all $n \in \mathbb{N}$, let $t_n := n\epsilon$. Moreover, let $\hat{T}_p^{(\epsilon)}$ be the smallest positive integer such that $(G_t)_{t\geq 0}$ jumps at least twice on the time interval $[t_{\hat{T}_p^{(\epsilon)}}, t_{\hat{T}_p^{(\epsilon)}+1})$, and let $\hat{T}_N^{(\epsilon)}$ be the smallest positive integer such that there exists $t \in [t_{\hat{T}_N^{(\epsilon)}}, t_{\hat{T}_N^{(\epsilon)}+1})$ such that $M_t \geq N$. We then construct the coupled discretized 2-CGP $(\hat{G}_n^{N,\epsilon})_{n\in\mathbb{N}}$ as follows.

1. First, we set $\hat{G}_0^{(N,\epsilon)} = 0$.
2. For all $n \in [\![0, \min(\hat{T}_N^{(\epsilon)}, \hat{T}_p^{(\epsilon)}) - 1]\!]$, we set
$$\hat{G}_{n+1}^{(N,\epsilon)} = M_{t_{n+1}} - m_{t_{n+1}}.$$
3. If $\hat{T}_p^{(\epsilon)} \leq \hat{T}_N^{(\epsilon)}$, $\hat{G}_{\hat{T}_p^{(\epsilon)}+1}^{(N,\epsilon)}$ is taken equal to the value of $M_t - m_t$ after the first jump of $(G_t)_{t\geq 0}$ over the time interval $[t_{\hat{T}_p^{(\epsilon)}}, t_{\hat{T}_p^{(\epsilon)}+1})$. Otherwise, we set $\hat{G}_{\hat{T}_N^{(\epsilon)}+1}^{(N,\epsilon)} = M_{t_{\hat{T}_N^{(\epsilon)}+1}} - m_{t_{\hat{T}_N^{(\epsilon)}+1}}$.
4. For $n > \min\left(\hat{T}_p^{(\epsilon)} + 1, \hat{T}_N^{(\epsilon)} + 1\right)$, the coupling no longer holds, and $(\hat{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ evolves according to the dynamics described in Definition 4.5.10.

This coupling satisfies the following property.

Lemma 4.5.11. *Let $\epsilon > 0$ and $N \in \mathbb{N}\backslash\{0,1\}$. Then, the coupling of the discretized 2-CGP $(\hat{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ with timestep ϵ and maximal height difference N to the original 2-CGP $(G_t)_{t\geq 0}$ holds until time $\min\left(\hat{T}_p^{(\epsilon)}, \hat{T}_N^{(\epsilon)}\right)$. In other words, for all $n \in \mathbb{N}$, if $n \leq \min\left(\hat{T}_p^{(\epsilon)}, \hat{T}_N^{(\epsilon)}\right)$, then*
$$\hat{G}_n^{(N,\epsilon)} = M_{t_n} - m_{t_n}.$$
Moreover, let $\hat{T}_\square^{(N,\epsilon)}$ be defined as
$$\hat{T}_\square^{(N,\epsilon)} := \min\left\{n \in \mathbb{N}\backslash\{0\} : \hat{G}_n^{(N,\epsilon)} = 0 \text{ but } \hat{G}_{n-1}^{(N,\epsilon)} \neq 0\right\}.$$
If $\hat{T}_\square^{(N,\epsilon)} \leq \min\left(\hat{T}_N^{(\epsilon)}, \hat{T}_p^{(\epsilon)}\right)$, then $\epsilon\hat{T}_\square^{(N,\epsilon)} - \epsilon < T_\square \leq \epsilon\hat{T}_\square^{(N,\epsilon)}$ a.s.

Notice that contrary to $T_\square$, the random variable $\hat{T}_\square^{(N,\epsilon)}$ does not correspond exactly to the time of first return of $\hat{G}_n^{(N,\epsilon)}$ to 0, but rather to the time needed for the process to exit state 0 and then return to it. For instance, if $\hat{G}_1^{(N,\epsilon)} = 0$, then the time of first return to state 0 of $(\hat{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ is equal to 1, while $\hat{T}_\square^{(N,\epsilon)} > 1$.

Proof. The first part of the lemma is a direct consequence of the coupling. Then, we assume that $\hat{T}_\square^{(N,\epsilon)} \leq \min\left(\hat{T}_N^{(\epsilon)}, \hat{T}_p^{(\epsilon)}\right)$. By definition of $\hat{T}_\square^{(N,\epsilon)}$, we know that
$$M_{t_{\hat{T}_\square^{(N,\epsilon)}}} - m_{t_{\hat{T}_\square^{(N,\epsilon)}}} = 0, \tag{4.5.9}$$
and for all $n \in [\![0, \hat{T}_\square^{(N,\epsilon)}]\!]$,
$$M_{t_n} - m_{t_n} \neq 0.$$

Moreover, $(G_t)_{t\geq 0}$ jumps at most once over each time interval $[t_n, t_{n+1})$, $n \in [\![0, \hat{T}_{\square}^{(N,\epsilon)} - 1]\!]$. Therefore, for all $n \in [\![0, \hat{T}_{\square}^{(N,\epsilon)} - 1]\!]$ and $t \in [t_n, t_{n+1})$,

$$M_t - m_t \in \left\{M_{t_n} - m_{t_n}, M_{t_{n+1}} - m_{t_{n+1}}\right\},$$

from which we deduce

$$\forall t \in \left[0, t_{\hat{T}_{\square}^{(N,\epsilon)}-1}\right], \quad M_t - m_t \neq 0.$$

Therefore,

$$T_{\square} > t_{\hat{T}_{\square}^{(N,\epsilon)}-1} = \epsilon\left(\hat{T}_{\square}^{(N,\epsilon)} - 1\right).$$

Moreover, by Eq.(4.5.9),

$$T_{\square} \leq t_{\hat{T}_{\square}^{(N,\epsilon)}} = \epsilon \hat{T}_{\square}^{(N,\epsilon)},$$

and we can conclude. □

The interest of the discretized 2-CGP lies in the fact that it is possible to compute explicitly $\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}]$, using the invariant distribution of the process. $\epsilon\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}]$ is a good approximation to $\mathbb{E}_{(0,0)}[T_{\square}]$ when $\epsilon \to 0$ and $N \to +\infty$, as stated in the following result.

Proposition 4.5.12. *We have*

$$\lim_{\substack{\epsilon\to 0 \\ N\to+\infty \\ N^2\epsilon\to 0}} \epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_{\square}^{(N,\epsilon)}\right] = \mathbb{E}_{(0,0)}[T_{\square}].$$

In order to show Proposition 4.5.12, we need two technical lemmas. The first one, Lemma 4.5.13, states that for large N, with high probability, the 2-CGP comes back to a state in $\mathcal{S}_{\square\square}$ before reaching height N. The second one, Lemma 4.5.14, states that before coming back to a state in $\mathcal{S}_{\square\square}$, under suitable conditions on ϵ and N, with high probability, at most one growth event occurs in each interval of length ϵ. Notice that these two lemmas describe the properties of the *original* (non-discretized) 2-CGP. We will use them to show that in the limiting regime, the coupling between the discretized 2-CGP and the original process still holds at time $\hat{T}_{\square}^{(N,\epsilon)}$, as whether the coupling breaks depends on the dynamics of the original 2-CGP.

Lemma 4.5.13. *For the non-discretized 2-CGP $(G_t)_{t\geq 0} = (m_t, M_t)_{t\geq 0}$, for all $\epsilon > 0$ and $N \in \mathbb{N}\backslash\{0,1\}$,*

$$\mathbb{P}_{(0,0)}\left(T_N \leq T_{\square} + \epsilon\right) \leq \frac{1}{2^{N-1}}\exp(\epsilon),$$

where we recall that $T_N = \inf\{t \geq 0 : M_t \geq N\}$.

Proof. Let $\epsilon > 0$ and $N \in \mathbb{N}\backslash\{0,1\}$. Adapting the proof of Lemma 4.5.7, we obtain that

$$\mathbb{P}_{(0,0)}\left(T_N \leq T_{\square} + \epsilon\right) \leq \mathbb{P}_{(0,0)}\left(\mathcal{P}\left(T_{\rightarrow} + \epsilon\right) \geq N - 1\right),$$

where $T_{\rightarrow} \sim \mathcal{E}xp(1)$ and where $\mathcal{P}(\lambda)$, $\lambda > 0$ stands for a Poisson random variable with parameter λ. Therefore,

$$\begin{aligned}\mathbb{P}_{(0,0)}\left(T_N \leq T_{\square} + \epsilon\right) &\leq \int_0^{\infty} e^{-t}e^{-(t+\epsilon)}\left(\sum_{i=N-1}^{+\infty}\frac{(t+\epsilon)^i}{i!}\right)dt \\ &= e^{-\epsilon}\left[\sum_{i=N-1}^{+\infty}\int_0^{\infty} e^{-2t}\,\frac{(t+\epsilon)^i}{i!}dt\right].\end{aligned}$$

Moreover, for all $i \geq N-1$,

$$\begin{aligned}\int_0^\infty e^{-2t}\frac{(t+\epsilon)^i}{i!}dt &= \left[-\frac{1}{2}e^{-2t}\frac{(t+\epsilon)^i}{i!}\right]_0^\infty + \int_0^\infty \frac{1}{2}e^{-2t}\frac{(t+\epsilon)^{i-1}}{(i-1)!}dt \\ &= \frac{1}{2}\frac{\epsilon^i}{i!} + \frac{1}{2}\int_0^\infty e^{-2t}\frac{(t+\epsilon)^{i-1}}{(i-1)!}dt \\ &= \sum_{j=0}^{i}\frac{1}{2^{j+1}}\frac{\epsilon^{i-j}}{(i-j)!} \\ &= \frac{1}{2^{i+1}}\left(\sum_{j=0}^{i}\frac{(2\epsilon)^{i-j}}{(i-j)!}\right) \\ &\leq \frac{1}{2^{i+1}}\exp(2\epsilon).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}_{(0,0)}\left(T_N \leq T_\square + \epsilon\right) &\leq \sum_{i=N-1}^{+\infty}\frac{1}{2^{i+1}}e^{2\epsilon}e^{-\epsilon} \\ &\leq e^{\epsilon}\frac{1}{2^{N-1}},\end{aligned}$$

which concludes the proof. □

Lemma 4.5.14. *In the notation of Lemma 4.5.13,*

$$\mathbb{P}_{(0,0)}\left(\hat{T}_p^{(\epsilon)} < \frac{T_\square}{\epsilon}\middle| T_N > T_\square + \epsilon\right) \leq 2\left(1-\exp(-(N+2)\epsilon)\right) + \frac{1}{1-\exp(-\epsilon)}\left(1-\exp(-(N+2)\epsilon)\right)^2.$$

Proof. We consider the time intervals $[t_0, t_1)$, $[t_1, t_2)$, ..., $[t_{\lfloor T_1\epsilon^{-1}\rfloor}, t_{\lfloor T_1\epsilon^{-1}\rfloor+1})$, ..., $[t_{\lfloor T_\square\epsilon^{-1}\rfloor}, t_{\lfloor T_\square\epsilon^{-1}\rfloor+1})$. We work conditional on $T_N > T_\square + \epsilon$. In order to have $\hat{T}_p^{(\epsilon)} < T_\square\epsilon^{-1}$, one of the following events need to occur.

1. Another growth event occurs in the time interval $[t_{\lfloor T_1\epsilon^{-1}\rfloor}, t_{\lfloor T_1\epsilon^{-1}\rfloor+1})$.
   This occurs with probability bounded from above by $1-\exp(-(N+2)\epsilon)$.
2. At least two growth events occur in at least one of the time intervals $[t_{\lfloor T_1\epsilon^{-1}\rfloor+1}, t_{\lfloor T_1\epsilon^{-1}\rfloor+2})$, ..., $[t_{\lfloor T_\square\epsilon^{-1}\rfloor-1}, t_{\lfloor T_\square\epsilon^{-1}\rfloor})$.
   For each time interval, this occurs with probability bounded from above by $(1-\exp(-(N+2)\epsilon))^2$.
3. Another growth event occurs in the time interval $[t_{\lfloor T_\square\epsilon^{-1}\rfloor}, t_{\lfloor T_\square\epsilon^{-1}\rfloor+1})$.
   Again, this occurs with probability bounded from above by $1-\exp(-(N+2)\epsilon)$.

Moreover, if only one growth event occurs during the time interval $[t_{\lfloor T_1\epsilon^{-1}\rfloor}, t_{\lfloor T_1\epsilon^{-1}\rfloor+1})$, then the random variable $\lfloor T_\square\epsilon^{-1}\rfloor - \lfloor T_1\epsilon^{-1}\rfloor$ is bounded from above by a geometric law with probability of success $1-\exp(-\epsilon)$. Therefore,

$$\begin{aligned}&\mathbb{P}_{(0,0)}\left(\hat{T}_p^{(\epsilon)} < \frac{T_\square}{\epsilon}\middle| T_N > T_\square + \epsilon\right) \\ &\leq 2(1-\exp(-(N+2)\epsilon)) + \sum_{k=2}^{+\infty}k\exp(-\epsilon)^{k-1}(1-\exp(-\epsilon))(1-\exp(-(N+2)\epsilon))^2 \\ &\leq 2\left(1-\exp(-(N+2)\epsilon)\right) + \frac{1}{1-\exp(-\epsilon)}\left(1-\exp(-(N+2)dt)\right)^2,\end{aligned}$$

and we can conclude. □

We can now show Proposition 4.5.12.

Proof. (Proposition 4.5.12) By Lemma 4.5.11, if $T_\square \leq \min(\hat{T}_N^{(\epsilon)}, \epsilon\hat{T}_p^{\epsilon})$, then

$$\epsilon\hat{T}_\square^{(N,\epsilon)} - \epsilon < T_\square \leq \epsilon\hat{T}_\square^{(N,\epsilon)}.$$

In particular, this is true if $T_N - \epsilon > T_\square$ and $\hat{T}_p^{(\epsilon)} \geq T_\square\epsilon^{-1}$. Then, by case disjunction,

$$\begin{aligned}\mathbb{E}_{(0,0)}[T_\square] =&\mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N<T_\square+\epsilon\}}\right] + \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)<T_\square\epsilon^{-1}\}}\right]\\ &+ \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)\geq T_\square\epsilon^{-1}\}}\right]\end{aligned}$$

and similarly

$$\begin{aligned}\mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\right] =&\mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\mathbb{1}_{\{T_N<T_\square+\epsilon\}}\right] + \mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)<T_\square\epsilon^{-1}\}}\right]\\ &+ \mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)\geq T_\square\epsilon^{-1}\}}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}&\epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)\geq T_\square\epsilon^{-1}\}}\right] - \epsilon\\ &\quad+ \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N<T_\square+\epsilon\}}\right] + \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)<T_\square\epsilon^{-1}\}}\right]\\ \leq&\mathbb{E}_{(0,0)}[T_\square]\\ \leq&\epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)\geq T_\square\epsilon^{-1}\}}\right]\\ &\quad+ \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N<T_\square+\epsilon\}}\right] + \mathbb{E}_{(0,0)}\left[T_\square\mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)<T_\square\epsilon^{-1}\}}\right],\end{aligned}$$

and hence

$$\begin{aligned}&\epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\right] - \epsilon + \mathbb{E}_{(0,0)}\left[\left(T_\square - \epsilon\hat{T}_\square^{N,\epsilon}\right)\left(\mathbb{1}_{\{T_N<T_\square+\epsilon\}} + \mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)<T_\square\epsilon^{-1}\}}\right)\right]\\ \leq&\mathbb{E}_{(0,0)}[T_\square]\\ \leq&\epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_\square^{(N,\epsilon)}\right] + \mathbb{E}_{(0,0)}\left[\left(T_\square - \epsilon\hat{T}_\square^{N,\epsilon}\right)\left(\mathbb{1}_{\{T_N<T_\square+\epsilon\}} + \mathbb{1}_{\{T_N>T_\square+\epsilon\}}\mathbb{1}_{\{\hat{T}_p(\epsilon)<T_\square\epsilon^{-1}\}}\right)\right].\end{aligned}$$

By Lemma 4.5.3, $\mathbb{E}_{(0,0)}[T_\square] \leq 3/2$. Moreover,

$$\begin{aligned}\mathbb{E}_{(0,0)}\left[\epsilon\hat{T}_\square^{(N,\epsilon)}\right] &\leq \epsilon\left(\frac{1}{\hat{p}_{0,0}^{(N,\epsilon)}} + \frac{1}{\hat{p}_{1,0}^{(N,\epsilon)}} + \max_{i\in[\![2,N]\!]}\frac{1}{\hat{p}_{i,0}^{(N,\epsilon)}}\right)\\ &= \epsilon\left(\frac{1}{1-\exp(-2\epsilon)} + \frac{3}{2(1-\exp(-3\epsilon)} + \max_{i\in[\![2,N]\!]}\frac{i+2}{1-\exp(-(i+2)\epsilon)}\right)\\ &\leq \epsilon\left(\frac{1}{1-\exp(-2\epsilon)} + \frac{3}{2(1-\exp(-3\epsilon)} + \frac{N+2}{1-\exp(-(N+2)\epsilon)}\right).\end{aligned}$$

Therefore,

$$\mathbb{E}_{(0,0)}\left[\left|T_\square - \epsilon\hat{T}_\square^{(N,\epsilon)}\right|\right] \leq 3/2 + \epsilon\left(\frac{1}{1-\exp(-2\epsilon)} + \frac{3}{2(1-\exp(-3\epsilon)} + \frac{N+2}{1-\exp(-(N+2)\epsilon)}\right)$$

and

$$\lim_{\substack{\epsilon\to 0\\ N\to+\infty\\ N^2\epsilon\to 0}} \mathbb{E}_{(0,0)}\left[\left|T_{\square} - \hat{T}_{\square}^{(N,\epsilon)}\right|\right] \leq 3/2 + 1/2 + 3/6 + 1 < +\infty.$$

Moreover, by Lemmas 4.5.13 and 4.5.14,

$$\mathbb{P}_{(0,0)}\left(T_N < T_{\square} + \epsilon\right) \xrightarrow[\substack{\epsilon\to 0\\ N\to+\infty\\ N^2\epsilon\to 0}]{} 0$$

$$\text{and } \mathbb{P}_{(0,0)}\left(\hat{T}_p^{\epsilon} < T_{\square}\epsilon^{-1} \middle| T_N > T_{\square} + \epsilon\right) \xrightarrow[\substack{\epsilon\to 0\\ N\to+\infty\\ N^2\epsilon\to 0}]{} 0,$$

which allows us to conclude using the dominated convergence theorem. □

Therefore, if we compute $\epsilon\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}]$ for N large enough and ϵ small enough, we can use the corresponding value to approximate $\mathbb{E}_{(0,0)}[T_{\square}]$. The next section is devoted to obtaining an explicit expression for $\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}]$, using the invariant distribution of $(\hat{G}_n^{(N,\epsilon)})_{n\geq 0}$.

4.5.3 Invariant distribution of the discretized 2-CGP

Since the discretized 2-CGP is an irreductible and positive recurrent Markov chain, there exists a relation between its invariant distribution and the expected first return times for each of its states. We want to use this relation to obtain an explicit expression for $\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}]$. However, as explained earlier, the definition of $\hat{T}_{\square}^{(N,\epsilon)}$ does not correspond to how first return times for discrete-time Markov chains are usually defined in the literature. Therefore, the invariant distribution of $(\hat{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ does not directly give access to $\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}]$.

In order to circumvent this problem, we now introduce the *accelerated discretized 2-CGP*, denoted $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$. The dynamics of this new process is identical to the one of the original discretized 2-CGP, *except when the process is in state* 0. In this case, the accelerated discretized 2-CGP jumps to state 1 with probability 1. Therefore, the process cannot stay in state 0 during more than one time step, and the time needed to first leave state 0, and then return to it is given by the invariant distribution of the process.

Definition 4.5.15. *The accelerated discretized 2-CGP $(\widetilde{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ with timestep ϵ and maximal height difference N is the $[\![0,N]\!]$-valued discrete-time Markov chain with initial condition $\widetilde{G}_0^{(N,\epsilon)} = 0$ and whose transition probabilities $(p_{i,j}^{(N,\epsilon)})_{(i,j)\in[\![0,N]\!]^2}$ are defined as follows.*

1. *If $i = 0$, then $p_{0,0}^{(N,\epsilon)} = 0$, $p_{0,1}^{(N,\epsilon)} = 1$, and for all $j \in [\![2,N]\!]$, $p_{0,j}^{(N,\epsilon)} = 0$.*
2. *For all $i \in [\![1,N]\!]$ and for all $j \in [\![0,N]\!]$, $p_{i,j}^{(N,\epsilon)} = \hat{p}_{i,j}^{(N,\epsilon)}$.*

Similarly as before, we will say that $(\tilde{G}_n^{(N,\epsilon)})_{n\in\mathbb{N}}$ starts from the state $(0,0)$ (resp. comes back to a state in $\mathcal{S}_{\square\square}$) when it starts from (resp. comes back to) the state 0.

As stated before, the main difference between $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$ and $(\widehat{G}_n^{(N,\epsilon)})_{n\geq 0}$ lies in the fact that $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$ cannot stay in the state 0 during more than one time step. Therefore, the mean time of first return to state 0 *starting from state* 1 are equal for the original process and its accelerated version. Moreover, we can compute this time for the accelerated discretized 2-CGP, since:

- Its mean time of first return to state 0 *starting from state* 0 can be computed using its invariant distribution.

- If it starts from state 0, then it reaches state 1 in exactly one timestep.

Therefore, if $\widetilde{T}_{\square}^{(N,\epsilon)}$ stands for the time of first return of $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$ to the state 0, we have the following lemma.

Lemma 4.5.16.

$$\mathbb{E}_{(0,0)}[\hat{T}_{\square}^{(N,\epsilon)}] = \mathbb{E}_{(0,0)}[\widetilde{T}_{\square}^{(N,\epsilon)}] - 1 + \frac{1}{1-\exp(-2\epsilon)}.$$

Proof. Indeed, $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$ exits the state 0 in exactly one time step, while $(\widehat{G}_n^{(N,\epsilon)})_{n\geq 0}$ needs a number of time steps distributed as a geometrical law with probability of success $1-\exp(-2\epsilon)$ to do so. □

We now compute the invariant distribution of $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$. In order to do so, let $(\tilde{\mathbf{p}}_i^{(N,\epsilon)})_{0\leq i\leq N}$ stand for the invariant distribution of $(\widetilde{G}_n^{(N,\epsilon)})_{n\geq 0}$. Let $(A_i^{(N,\epsilon)})_{0\leq i\leq N}$ be defined by backwards induction as follows.

1. First, we set $A_N^{(N,\epsilon)} = 1$, $A_{N-1}^{(N,\epsilon)} = N+1$ and $A_{N-2}^{(N,\epsilon)} = (N+1)A_{N-1}^{(N,\epsilon)} - 2A_N^{(N,\epsilon)}$.
2. Then, for all $i \in [\![2, N-2]\!]$, we set

$$A_{i-1}^{(N,\epsilon)} = (i+2)A_i^{(N,\epsilon)} - 2A_{i+1}^{(N,\epsilon)} - \sum_{j=i+2}^{N} A_j^{(N,\epsilon)}.$$

3. We conclude by setting

$$A_0^{(N,\epsilon)} = \frac{1}{2}\left(2A_1^{(N,\epsilon)} + \sum_{j=2}^{N} A_j^{(N,\epsilon)}\right)(1-\exp(-2\epsilon)).$$

Then, the sequence $(\tilde{\mathbf{p}}_i^{(N,\epsilon)})_{0\leq i\leq N}$ can be expressed in terms of the sequence $(A_i^{(N,\epsilon)})_{0\leq i\leq N}$ as follows.

Lemma 4.5.17. *For all $i \in [\![0, N]\!]$,*

$$\tilde{\mathbf{p}}_i^{(N,\epsilon)} = \tilde{\mathbf{p}}_N^{(N,\epsilon)} A_i^{(N,\epsilon)} \frac{i+2}{N+2} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(i+2)\epsilon)}.$$

Proof. We show that the result is true by backwards induction. First, we check that it is true for $i = N$, $i = N-1$ and $i = N-2$.

$$A_N^{(N,\epsilon)} \frac{N+2}{N+2} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(N+2)\epsilon)} = A_N^{(N,\epsilon)} = 1,$$

so the property is true for $i = N$. Then, by definition of the invariant distribution,

$$\tilde{\mathbf{p}}_{N-1}^{(N,\epsilon)} p_{N-1,N}^{(N,\epsilon)} + \tilde{\mathbf{p}}_N^{(N,\epsilon)} p_{N,N}^{(N,\epsilon)} = \tilde{\mathbf{p}}_N^{(N,\epsilon)}$$

and so

$$\begin{aligned}
\tilde{\mathbf{p}}_{N-1}^{(N,\epsilon)} &= \frac{1}{p_{N-1,N}^{(N,\epsilon)}} \tilde{\mathbf{p}}_N^{(N,\epsilon)} (1 - p_{N,N}^{(N,\epsilon)}) \\
&= \frac{N+1}{1-\exp(-(N+1)\epsilon)} \tilde{\mathbf{p}}_N^{(N,\epsilon)} \left(\frac{N+1}{N+2}(1-\exp(-(N+2)\epsilon))\right) \\
&= \tilde{\mathbf{p}}_N^{(N,\epsilon)} \frac{N+1}{N+2}(N+1) \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(N+1)\epsilon)}
\end{aligned}$$

and the result is true for $i = N-1$. Moreover,

$$\tilde{\mathbf{p}}_{N-2}^{(N,\epsilon)} p_{N-2,N-1}^{(N,\epsilon)} + \tilde{\mathbf{p}}_{N-1}^{(N,\epsilon)} p_{N-1,N-1}^{(N,\epsilon)} + \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} p_{N,N-1}^{(N,\epsilon)} = \tilde{\mathbf{p}}_{N-1}^{(N,\epsilon)},$$

which means that

$$\begin{aligned}
\tilde{\mathbf{p}}_{N-2}^{(N,\epsilon)} &= \frac{1}{p_{N-2,N-1}^{(N,\epsilon)}} \left[\tilde{\mathbf{p}}_{N-1}^{(N,\epsilon)} (1 - p_{N-1,N-1}^{(N,\epsilon)}) - \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} p_{N,N-1}^{(N,\epsilon)} \right] \\
&= \frac{N}{1-\exp(-N\epsilon)} \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} \left[\frac{N+1}{N+2} A_{N-1}^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(N+1)\epsilon)} \left(1-\exp(-(N+1)\epsilon)\right) \right] \\
&\quad - \frac{N}{1-\exp(-N\epsilon)} \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} \left[\frac{2}{N+2} \left(1-\exp(-(N+2)\epsilon)\right) \right] \\
&= \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} \frac{N}{N+2} \left((N+1) A_{N-1}^{(N,\epsilon)} - 2 A_{N}^{(N,\epsilon)} \right) \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-N\epsilon)} \\
&= \tilde{\mathbf{p}}_{N} \frac{N}{N+2} A_{N-2}^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-N\epsilon)}
\end{aligned}$$

by definition of $A_{N-2}^{(N,\epsilon)}$.

Then, let $i \in [\![2, N-2]\!]$, and assume that the property is true for $j \in [\![i, N]\!]$. Again by definition of the invariant property,

$$\tilde{\mathbf{p}}_{i-1}^{(N,\epsilon)} p_{i-1,i}^{(N,\epsilon)} + \tilde{\mathbf{p}}_{i}^{(N,\epsilon)} p_{i,i}^{(N,\epsilon)} + \tilde{\mathbf{p}}_{i+1}^{(N,\epsilon)} p_{i+1,i}^{(N,\epsilon)} + \sum_{j=i+2}^{N} \tilde{\mathbf{p}}_{j}^{(N,\epsilon)} p_{j,i}^{(N,\epsilon)} = \tilde{\mathbf{p}}_{i}^{(N,\epsilon)},$$

from which we deduce

$$\begin{aligned}
\tilde{\mathbf{p}}_{i-1}^{(N,\epsilon)} &= \frac{1}{p_{i-1,i}^{(N,\epsilon)}} \left(\tilde{\mathbf{p}}_{i}^{(N,\epsilon)} (1 - p_{i,i}^{(N,\epsilon)}) - \tilde{\mathbf{p}}_{i+1}^{(N,\epsilon)} p_{i+1,i}^{(N,\epsilon)} - \sum_{j=i+2}^{N} \tilde{\mathbf{p}}_{j}^{(N,\epsilon)} p_{j,i}^{(N,\epsilon)} \right) \\
&= \frac{i+1}{1-\exp(-(i+1)\epsilon)} \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} \\
&\quad \times \left[\frac{i+2}{N+2} A_{i}^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(i+2)\epsilon)} (1-\exp(-(i+2)\epsilon)) \right. \\
&\qquad - \frac{i+3}{N+2} A_{i+1}^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(i+3)\epsilon)} \frac{2}{i+3} (1-\exp(-(i+3)\epsilon)) \\
&\qquad \left. - \sum_{j=i+2}^{N} \frac{j+2}{N+2} A_{j}^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(j+2)\epsilon)} \frac{1}{j+2} (1-\exp(-(j+2)\epsilon) \right] \\
&= \tilde{\mathbf{p}}_{N}^{(N,\epsilon)} \frac{i+1}{N+2} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(i+1)\epsilon)} \left((i+2) A_{i}^{(N,\epsilon)} - 2 A_{i+1}^{(N,\epsilon)} - \sum_{j=i+2}^{N} A_{j}^{(N,\epsilon)} \right) \\
&= \tilde{\mathbf{p}}_{N} A_{i-1}^{(N,\epsilon)} \frac{i+1}{N+2} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-(i+1)\epsilon)}
\end{aligned}$$

by definition of $A_{i-1}^{(N,\epsilon)}$.

We now need to check that the property is true for $i = 0$. We have

$$\tilde{\mathbf{p}}_{1}^{(N,\epsilon)} p_{1,0}^{(N,\epsilon)} + \sum_{j=2}^{N} \tilde{\mathbf{p}}_{j}^{(N,\epsilon)} p_{j,0}^{(N,\epsilon)} = \tilde{\mathbf{p}}_{0}^{(N,\epsilon)},$$

so

$$\begin{aligned}
\tilde{\mathbf{p}}_0^{(N,\epsilon)} &= \left[\tilde{\mathbf{p}}_1^{(N,\epsilon)} p_{1,0}^{(N,\epsilon)} + \sum_{j=2}^{N} \tilde{\mathbf{p}}_j^{(N,\epsilon)} p_{j,0}^{(N,\epsilon)}\right] \\
&= \tilde{\mathbf{p}}_N^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{N+2} \\
&\quad \times \left[3\frac{2}{3}\frac{1-\exp(-3\epsilon)}{1-\exp(-3\epsilon)} A_1^{(N,\epsilon)} + \sum_{j=2}^{N} \frac{(j+2)A_j^{(N,\epsilon)}}{1-\exp(-(j+2)\epsilon)} \frac{1}{j+2}(1-\exp(-(j+2)\epsilon))\right] \\
&= \tilde{\mathbf{p}}_N^{(N,\epsilon)} \frac{1}{N+2} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-2\epsilon)} \left(2A_1^{(N,\epsilon)} + \sum_{j=2}^{N} A_j^{(N,\epsilon)}\right)(1-\exp(-2\epsilon)) \\
&= \tilde{\mathbf{p}}_N^{(N,\epsilon)} \frac{2}{N+2} A_0^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{1-\exp(-2\epsilon)},
\end{aligned}$$

which allows us to conclude. □

We can now use the invariant distribution to obtain an explicit formula for $\mathbb{E}_{(0,0)}\left[\hat{T}_{\square}^{(N,\epsilon)}\right]$.

Proposition 4.5.18.

$$\mathbb{E}_{(0,0)}\left[\hat{T}_{\square}^{(N,\epsilon)}\right] = \frac{1}{1-\exp(-2\epsilon)} + \frac{1-\exp(-2\epsilon)}{2A_0^{(N,\epsilon)}} \left(\sum_{i=1}^{N} \frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right).$$

Proof. We know that

$$\mathbb{E}_{(0,0)}\left[\widetilde{T}_{\square}^{(N,\epsilon)}\right] = \frac{1}{\tilde{\mathbf{p}}_0^{(N,\epsilon)}}$$

and

$$\sum_{i=0}^{N} \tilde{\mathbf{p}}_i^{(N,\epsilon)} = 1.$$

Using Lemma 4.5.17, we obtain that

$$\begin{aligned}
\tilde{\mathbf{p}}_N^{(N,\epsilon)} \frac{1-\exp(-(N+2)\epsilon)}{N+2} \left[\sum_{i=0}^{N} \frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right] &= 1 \\
\text{and so } \tilde{\mathbf{p}}_N^{(N,\epsilon)} &= \frac{N+2}{1-\exp(-(N+2)\epsilon)} \frac{1}{\sum_{i=0}^{N} \frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}}.
\end{aligned}$$

Using again Lemma 4.5.17 yields

$$\begin{aligned}
\tilde{\mathbf{p}}_0^{(N,\epsilon)} &= 2A_0^{(N,\epsilon)} \frac{1}{1-\exp(-2\epsilon)} \left(\sum_{i=0}^{N} \frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right)^{-1} \\
&= \frac{1}{1+\left[\sum_{i=1}^{N} \frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right] \frac{1-\exp(-2\epsilon)}{2A_0^{(N,\epsilon)}}},
\end{aligned}$$

and using Lemma 4.5.16, we obtain

$$\begin{aligned}\mathbb{E}_{(0,0)}\left[\hat{T}_{\square}^{(N,\epsilon)}\right] &= \mathbb{E}_{(0,0)}\left[\widetilde{T}_{\square}^{(N,\epsilon)}\right] - 1 + \frac{1}{1-\exp(-2\epsilon)} \\ &= 1 + \frac{1-\exp(-2\epsilon)}{2A_0^{(N,\epsilon)}}\left(\sum_{i=1}^{N}\frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right) - 1 + \frac{1}{1-\exp(-2\epsilon)} \\ &= \frac{1}{1-\exp(-2\epsilon)} + \frac{1-\exp(-2\epsilon)}{2A_0^{(N,\epsilon)}}\left(\sum_{i=1}^{N}\frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right).\end{aligned}$$

□

We can now use Proposition 4.5.12 in order to compute an approximation for $\mathbb{E}_{(0,0)}[T_{\square}]$.

Proposition 4.5.19.

$$\mathbb{E}_{(0,0)}[T_{\square}] = \frac{1}{2} + \lim_{\substack{\epsilon\to 0\\ N\to+\infty\\ N^2\epsilon\to 0}} \sum_{i=1}^{N}\left((i+2)\frac{A_i^{(N,\epsilon)}}{2A_0^{(N,\epsilon)}}\frac{1-\exp(-2\epsilon)}{1-\exp(-(i+2)\epsilon)}\right).$$

Proof. By Proposition 4.5.12,

$$\mathbb{E}_{(0,0)}[T_{\square}] = \lim_{\substack{\epsilon\to 0\\ N\to+\infty\\ N^2\epsilon\to 0}} \epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_{\square}^{(N,\epsilon)}\right].$$

Moreover, by Proposition 4.5.18,

$$\epsilon\mathbb{E}_{(0,0)}\left[\hat{T}_{\square}^{(N,\epsilon)}\right] = \frac{\epsilon}{1-\exp(-2\epsilon)} + \frac{\epsilon(1-\exp(-2\epsilon))}{2A_0^{(N,\epsilon)}}\left(\sum_{i=1}^{N}\frac{(i+2)A_i^{(N,\epsilon)}}{1-\exp(-(i+2)\epsilon)}\right).$$

We conclude using the fact that

$$\frac{\epsilon}{1-\exp(-2\epsilon)} \xrightarrow[\epsilon\to 0]{} \frac{1}{2}.$$

□

We obtain that $\mathbb{E}_{(0,0)}[T_{\square}] \simeq 1.46$, which is higher than the lower bound of 1.33 obtained at the end of Section 4.5.1.

4.6 Appendix : Geometrical properties of ellipses

In this section, we show some geometrical properties of ellipses, which are used in other sections. In all that follows, let $z_c = (x_c, y_c) \in \mathbb{R}^2$, $(a,b) \in (0,+\infty)$ and $\gamma \in (-\pi/2, \pi/2)$. We recall that $\mathfrak{B}_{a,b,\gamma}(z)$ is the ellipse defined by:

$$\mathfrak{B}_{a,b,\gamma}(z_c) = \left\{\begin{pmatrix}x_c\\ y_c\end{pmatrix} + A_\gamma\begin{pmatrix}ar\cos(\theta)\\ br\sin(\theta)\end{pmatrix} : r\in[0,1], \theta\in[0,2\pi)\right\}$$

where

$$A_\gamma = \begin{pmatrix}\cos(\gamma) & -\sin(\gamma)\\ \sin(\gamma) & \cos(\gamma)\end{pmatrix}.$$

The first lemma gives the maximal *horizontal separation* between a point in the ellipse and its center. This result is used in Section 4.3 to construct the express chain.

Lemma 4.6.1. *Let $f : [0,1] \times [-\pi, \pi) \to \mathbb{R}$ be the function defined by*

$$\forall (r, \theta) \in [0,1] \times [-\pi, \pi), f(r, \theta) = ar\cos(\theta)\cos(\gamma) - br\sin(\theta)\sin(\gamma).$$

Then, f reaches its maximum for

$$(r_{max}, \theta_{\max}) = \left(1, \arctan\left(-\frac{b}{a}\tan(\gamma)\right)\right),$$

and

$$f(r_{max}, \theta_{max}) = \sqrt{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}.$$

Proof. First, $r_{max} = 1$. Moreover, $\cos(\theta_{max})$ is of the same parity as $\cos(\gamma)$, and $\sin(\theta)$ is of opposite parity from $\sin(\gamma)$. As $\gamma \in (-\pi/2, \pi/2)$, we obtain that $\theta_{\max} \in (-\pi/2, \pi/2)$.

The function $f_\theta : (-\pi/2, \pi/2) \to \mathbb{R}$ such that for all $\theta \in (-\pi/2, \pi/2)$,

$$f_\theta(\theta) := a\cos(\theta)\cos(\gamma) - b\sin(\theta)\sin(\gamma)$$

is differentiable, and for all $\theta \in (-\pi/2, \pi/2)$,

$$f'_\theta(\theta) = -a\sin(\theta)\sin(\gamma) - b\cos(\theta)\sin(\gamma).$$

Therefore,

$$\begin{aligned}
& & f'_\theta(\theta_{max}) &= 0 \\
\iff & & a\sin(\theta_{max})\cos(\gamma) &= -b\cos(\theta_{max})\sin(\gamma) \\
\iff & & \tan(\theta_{max}) &= -\frac{b}{a}\tan(\gamma) \\
\iff & & \theta_{max} &= \arctan\left(-\frac{b}{a}\tan(\gamma)\right).
\end{aligned}$$

Moreover, since $\cos(\gamma) > 0$,

$$\begin{aligned}
& a\cos(\gamma)\cos(\theta_{max}) - b\sin(\gamma)\sin(\theta_{max}) \\
&= a\cos(\gamma)\cos\left(\arctan\left(-\frac{b}{a}\tan(\gamma)\right)\right) - b\sin(\gamma)\sin\left(\arctan\left(-\frac{b}{a}\tan(\gamma)\right)\right) \\
&= a\cos(\gamma)\frac{1}{\sqrt{1+\frac{b^2}{a^2}\tan^2(\gamma)}} + b\sin(\gamma)\frac{\frac{b}{a}\tan(\gamma)}{\sqrt{1+\frac{b^2}{a^2}\tan^2(\gamma)}} \\
&= \frac{a^2\cos^2(\gamma)}{a\cos(\gamma)\sqrt{1+\frac{b^2}{a^2}\tan^2(\gamma)}} + \frac{1}{a\cos(\gamma)}\frac{b^2\sin^2(\gamma)}{\sqrt{1+\frac{b^2}{a^2}\tan^2(\gamma)}} \\
&= \frac{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}{\sqrt{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}} \\
&= \sqrt{a^2\cos^2(\gamma) + b^2\sin^2(\gamma)}.
\end{aligned}$$

□

The second lemma means that the mean horizontal separation of a point in the ellipse from its center is equal to 0. Therefore, when a point is affected by a reproduction event, the mean horizontal separation of the center of the corresponding ellipse from the point is equal to 0.

Lemma 4.6.2. *Let $Z = (x+X, y+Y)$ be sampled uniformly at random in the ellipse $\mathfrak{B}_{a,b,\gamma}(z)$. Then,*

$$\mathbb{E}[X] = 0.$$

The proof is a straightforward symmetry argument.

Part III

Seed banks and expanding populations in urban tree bases

Chapter 5

Detecting seed bank influence on plant metapopulation dynamics

This chapter is based on a joint work with Nathalie Machon, Jean-Baptiste Mihoub and Alexandre Robert [Lou+21].

This is the accepted version of the following article: "Detecting seed bank influence on plant metapopulation dynamics", Louvet, A., Machon, N., Mihoub, J. B., and Robert, A. (2021). Methods in Ecology and Evolution, 12(4), 655-664, which has been published in final form at

https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13547.

Abstract

1. Seed banks are known to play a key role in plant metapopulations. However, detecting seed banks remains challenging and requires intense monitoring efforts. Assessing the genuine effect of seed banks on plant metapopulation dynamics (rather than their presence) may offer a much easier while still biologically relevant way to overcome this issue.

2. In this study, we developed a new metric : the Seed Bank Characteristic Event (SBCE) probability. Instead of detecting seed bank directly, the SBCE probability measures seed bank contribution to the observed metapopulation dynamics. Exploring seed bank parameters (colonization, germination and seed bank death probabilities, initial proportion of patches containing a seed bank), a wide range of monitoring durations (from 3 to 10 years) and number of patches in the metapopulation (from 10 to 1000 patches), we examined the conditions under which the SBCE probability is correctly estimated. To test the robustness of our approach, we further introduced false negatives, false positives or parameter heterogeneity between patches. Finally, we applied the SBCE probability method to the monitoring of tree bases plant species in Paris, France, to assess the applicability of the method to real-world datasets and increase understanding of plant metapopulation dynamics within an urban environment.

3. Our results indicate that the SBCE probability is well estimated when enough monitoring years or number of patches are considered, and for probabilities of false negatives or false positives of up to 0.1. However, the SBCE probability estimation is not robust to colonization probability heterogeneity between patches. When we applied the SBCE probability method to the real monitoring dataset, we found a contrasted contribution of the seed bank to the observed metapopulation dynamics from one street and one species to another.

4. The study suggests that the measurement of seed bank contribution is less data-demanding than assessment of seed bank presence. Applying the estimation method to the monitoring of tree bases plant species highlights a significant contribution of the seed bank to plant metapopulation dynamics in an urban environment, and illustrates how the method can be applied on real-world datasets.

5.1 Introduction

An important issue in ecology and conservation biology is determining the mechanisms underlying persistence of plant or animal populations in fragmented landscapes [Fah03]. In plants, the seed bank, i.e the spontaneous storage of seeds within the soil, plays a critical role in metapopulation and community dynamics [Fen95]. However, assessing directly the presence of a seed bank, by putting soil samples in germination chambers or by using seed separation methods [BB14], and measuring its associated parameters is challenging. Therefore conceptual approaches and statistical tools allowing one to estimate these quantities using widespread data such as patch occupancy data can prove very useful. One suitable conceptual framework for studying patchy environments is the metapopulation theory, first introduced in [Lev69]. A metapopulation is defined as a population living in a set of patches that can be colonized or go extinct, the regional persistence of the species resulting from a balance between local colonizations and extinctions [MW67]. Statistical tools have been developed to allow parameter inference for a broad range of metapopulation models (see e.g [Moi99; Moi04]), and were fruitfully used in studies on insects [Han11; MSH98] or small mammals [Ozg+06].
As plants form populations with a strong spatial structure and can only move from one patch to another as propagules, metapopulation models appear at first as particularly suited to their study [HB96]. Yet classical metapopulation models do not account for seed banks, which are common in seed plants [BB14], potentially leading to erroneous estimates of extinction and colonization rates [Fré+13] and making these models generally irrelevant for studying plant metapopulation dynamics [FW02]. New models taking into account the influence of a seed bank were developed recently [Fré+13; Bor+15]. These models consider the seed bank state as an *hidden state*, which is not visible but which influences patch occupancy, and which can be estimated from patch occupancy data. In this chapter, we elaborated a new method to characterize seed bank contribution to metapopulation dynamics based on the model introduced in [Plu+18]. This model allows parameter inference on a variant with a seed bank of a classical model, the *Propagule Rain Model* (or PRM) [Got91], in which patches are colonized or go extinct independently from each other, with the same colonization or extinction probability. However, a limit of the model proposed by [Plu+18] is that it would not be applicable in many real-world situations. It indeed requires either a long monitoring duration or several thousand patches to be monitored in order to accurately estimate all parameters (i.e. germination, colonization and seed bank death probabilities). To overcome this problem, we introduced in the present study a new metric providing information about the influence of the seed bank (rather than the seed bank presence per se) on plant population dynamics in real populations. We called this metric the *seed bank characteristic event* probability (SBCE probability). As it only measures the contribution of the seed bank to the observed standing vegetation dynamics, it is less informative than knowing all seed bank parameters, but we aim to show that it can be used in more real-life situations. Then, we used this metric with both theoretical and real metapopulation data. We performed analyses based on simulated presence/absence time series data, and on time series data in which we introduced some flaws commonly found in real datasets, in order to study estimation accuracy.
As a case study, we used the estimation method on annual floristic inventories of natural and spontaneous flora carried out from 2009 to 2018 on 1324 tree bases located in Paris, France (the *Paris 12* dataset). Indeed, the population of plants in urban tree bases is located inside an inhospitable matrix and has an high turnover, which makes metapopulation models particularly suited [DPC11]. Studies using presence/absence data for various species present in urban tree bases considered as metapopulations were already carried out, but to our knowledge none accounted for seed bank potential presence in tree bases [Oma+19; DPC11]. We also attempted to relate the SBCE probability to species traits and environmental characteristics.
Overall, our study (i) gives insights on the importance of seed bank contribution to plant metapop-

ulation dynamics within an urban environment, and (ii) provides a comprehensive framework to detect the effects of seed banks in plant metapopulations, which can be applied on a wide range of ecological systems, including but not restricted to urban environments.

5.2 Material and methods

5.2.1 Model used

The model from [Plu+18] we used in this study is a variant of the Propagule Rain Model (PRM) [Got91]. In the PRM, colonization and extinction probabilities do not depend on the current state of the metapopulation, and are constant over patches and time. A seed bank can be introduced using Hidden Markov Model (HMM) techniques [CMR05; Rab89]. The seed bank contains the seeds that were just produced by standing vegetation of the focal patch or came from colonization events by the propagule rain, along with seeds produced by previous generations that did not germinate yet and are still alive. Since all plants originate from the seed bank of the patch they are in, the presence of plants at one time step means that the seed bank contained seeds just before germination could occur. Therefore metapopulation parameters can be estimated along with the presence/absence of seeds in the seed bank at each time-step.
The model is characterized by these three parameters :

- the joint probability of seed germination and of survival of seedlings until adulthood. This parameter will be called *germination probability*, and denoted g, following the existing litterature.
- the *colonization probability* c of the patch by external seeds entering the seed bank.
- the *seed bank extinction probability conditional on the seed bank not having germinated* d. This parameter is the probability that the seed bank will not survive until the next generation, assuming it has not germinated yet.

The initial proportion p_0 of patches containing a seed bank can be considered as an extra parameter of the model.
The model evolves as follows : for each patch, if the seed bank is not empty, seeds can germinate with probability g. If they do, the plants will grow and produce seeds which will refill the seed bank. Otherwise, the seed bank can survive until the next generation with probability $1-d$ (denoted s in [Plu+18]). New seeds can enter the seed bank during a colonization event with probability c independently of the presence of standing vegetation or seeds in any given patch (see Figure 5.1). Standing vegetation produces seeds with probability 1 instead of probability $p \leq 1$ because otherwise the model would not be identifiable.

The main difference with the PRM model is that if the seed bank death probability d is strictly less than 1, then germination can be delayed. Hereafter, models for which $d < 1$ will be denoted as SB+ (*Seed-Bank Plus*) and those for which $d = 1$, corresponding to the classical PRM, will be denoted SB- (*Seed-Bank Minus*).
The procedure for estimating parameters uses the Expectation-Maximization (EM) algorithm [DLR77] in order to find the best parameter fit (see details in [Plu+18]). We set the number of iterations in the algorithm to 200 in order to ensure convergence of parameters. As the EM algorithm can converge to a local but non-global maximum, the choice of the initial conditions can affect the value returned by the algorithm. We preliminary checked with simulated datasets that convergence of the EM algorithm to a non-global maximum was rather unlikely (results not shown). We also implemented a variant of the method supporting years of missing data, using standard methods for HMM (see Supporting Information A).

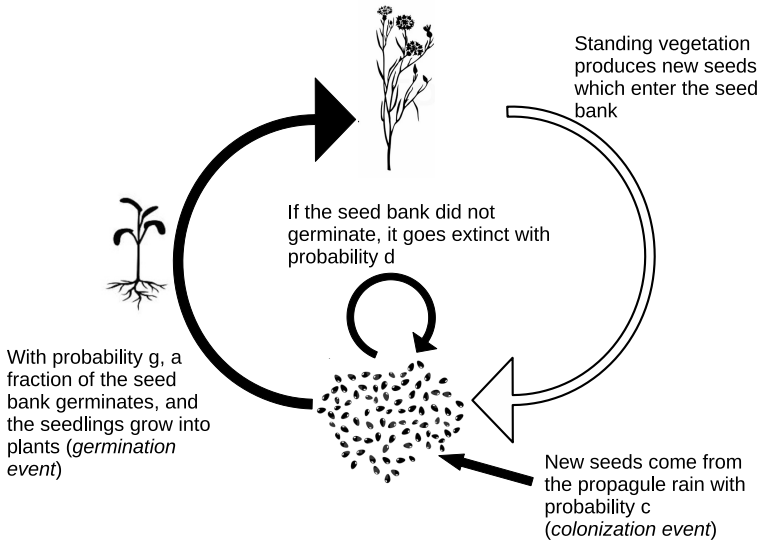
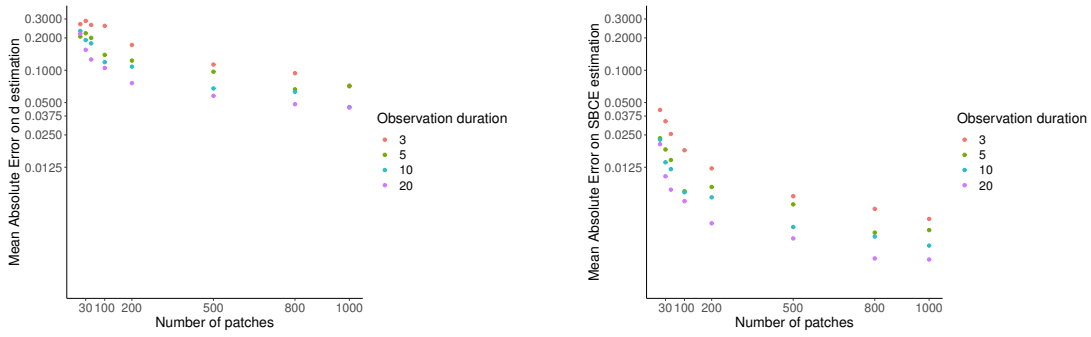


Figure 5.1: Graphical representation of the variant with seed bank of the Propagule Rain Model. White arrows indicate transitions that always occur, while black arrows indicate transitions that occur with a fixed probability.

As the estimator is asymptotically consistent, if enough patches and/or years of observation are considered, the estimated value for d will be equal or very close to 1 for models without seed bank, and different from 1 for models with seed bank. However, in many real-world situations, the estimation of d is not accurate enough to allow one to distinguish between models with $(d < 1)$ or without $(d = 1)$ seed bank (see Figure 5.2). Therefore, we introduced a new criterium for seed bank identification : the SBCE probability, which can be computed using g, c and d estimates.



(a) MAE on the estimation of d

(b) MAE on the estimation of $\mathbb{P}_{SBCE}$

Figure 5.2: Mean Absolute Error (MAE) on d estimation (graph (a)) and SBCE probability (graph (b)) estimation for SB- models (logarithmic scale). 30 datasets were generated for each combination of parameter values, number of patches and years of observation as presented in Table 5.1.

5.2.2 SBCE probability

Seed Bank Characteristic Event (SBCE) probability, denoted $\mathbb{P}_{SBCE}$, is defined as the probability for standing flora to produce seeds that (1) will stay in the seed bank during more than one year (without germinating nor dying), and (2) will germinate before new seeds come from an external source. Denoting SB_t the event *during t years, the seed bank does not germinate, nor receive new*

seeds from external source, nor dies, then :

$$\begin{aligned}\mathbb{P}_{SBCE} &= \sum_{t\geq 1} \mathbb{P}(SB_t) \times g \\ &= g \times \sum_{t\geq 1} \mathbb{P}(SB_1)^t \\ &= g \times \sum_{t\geq 1} [(1-g)(1-c)(1-d)]^t \\ &= g \times \frac{(1-g)(1-c)(1-d)}{1-(1-g)(1-c)(1-d)}\end{aligned}$$

To clarify the biological significance of SBCE probability, we considered two metapopulations of 100 patches each, evolving respectively under the SB+ and SB- models, with the same colonization and germination probabilities, and in which all 100 patches were initially occupied by plants. For each metapopulation, we counted the number of patches containing plants coming from seeds all produced by the plants initially present. This number is then on average equal to $100 \times g$ for the SB- model, and to $100 \times g + 100 \times \mathbb{P}_{SBCE}$ for the SB+ model.
In other words, $\mathbb{P}_{SBCE}$ is close to (but not equal to) the proportion of patches in the monitored metapopulation over time which contain plants, but would not in the absence of a seed bank. This metric will be called *seed bank proportional occupancy gain* thereafter, and measures the proportion of patches which would not be in the right observed state if the metapopulation was simulated without accounting for the seed bank. The relationship between $\mathbb{P}_{SBCE}$ and the seed bank proportional occupancy gain is illustrated in Figure 5.3 and in Section 5.5.2, and the method used to compute this metric is provided in Section 5.5.2. Importantly, $\mathbb{P}_{SBCE}$ is equal to zero in cases where there is no seed bank, but also in cases where there is a seed bank that does not contribute to the observed dynamics (i.e. when $c = 1$).

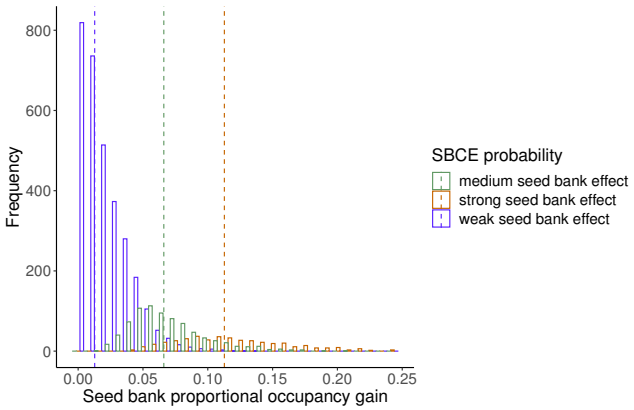


Figure 5.3: Relationship between the SBCE probability and the *seed bank proportional occupancy gain*. We modelled 4335 distinct parameter sets of large metapopulations over $t = 50$ consecutive years. For each parameter set, values of the seed bank proportional occupancy gain and the SBCE probability were computed. Parameter sets were then classified in three groups according to their SBCE probability value (weak, medium or strong seed bank effect). Dotted vertical lines indicate the median seed bank proportional occupancy gain for each group. The description of parameter sets and the method used to approximate the seed bank proportional occupancy gain are presented in Section 5.5.2.

Hereafter, we will consider the effect of the seed bank to be :

- *weak* if SBCE probability is lower than 0.05. In this case, both colonization and seed bank death probabilities are medium, or one of them is high. For metapopulations having such SBCE probability, the seed bank proportional occupancy gain is of at most $\sim 5\%$ (see Figure 5.3 and Section 5.5.2).
- *medium* if SBCE probability is in the interval $[0.05, 0.10]$. In this case, either the colonization of the seed bank death probability is low, but not both. For metapopulations having such SBCE probability, the seed bank proportional occupancy gain is around $5 - 10\%$.
- *strong* if SBCE probability is higher than 0.10. In this case, both colonization and seed bank death probabilities are low. For metapopulations having such SBCE probability, the seed bank proportional occupancy gain is above $\sim 10\%$.

An estimator of SBCE probability must satisfy two requirements so that it can accurately show seed bank contribution to metapopulation dynamics. First, the estimated SBCE probability must be accurately estimated for datasets generated with a SB+ model. Then, it must avoid, as much as possible, identification of medium or strong seed bank effects in datasets generated with a SB- model, for which $\mathbb{P}_{SBCE} = 0$. In other words, we will not try to distinguish models without seed bank from models with a seed bank which is not contributing significantly to the global metapopulation dynamics.
In order to assess the estimator accuracy, we investigated the sensibility of $\mathbb{P}_{SBCE}$ estimation to c, g, d and p_0 by building 24 distinct parameter sets combining a broad range of values for these parameters (16 corresponding to SB+ models, and 8 to SB- models, see Table 5.1). Each parameter set was used to generate 30 datasets for each of the 4 monitoring durations and 8 numbers of patches listed in Table 5.1. We chose to consider durations of at most 20 years and at most 1000 patches.
We computed the estimated SBCE probability by performing parameter fits of c, g, d and p_0 on the simulated datasets. The accuracy of SBCE probability estimation was tested separately for each SB+ parameter set, time duration and number of patches by computing the Mean Absolute Error (MAE). For SB- parameter sets, for each combination of monitoring duration and number of patches, we computed the proportion p_{fms} of datasets for which the estimated SBCE probability was above 0.05. We required this proportion to be below 0.05.

Table 5.1: Parameter sets and test conditions used in the study.

Parameter and notation	SB+ model	SB- model
Colonization probability (c)	0.3, 0.7	0.3, 0.7
Germination probability (g)	0.3, 0.7	0.3, 0.7
Seed bank death probability (d)	0.2, 0.6	1
Initial proportion of seeds (p_0)	0.3, 0.7	0.3, 0.7
Years of observation	3, 5, 10, 20	3, 5, 10, 20
Number of patches	10, 30, 50, 100, 200, 500, 800, 1000	10, 30, 50, 100, 200, 500, 800, 1000

5.2.3 Testing SBCE estimation robustness

Before applying the parameter estimation method to real data, we tested the algorithm robustness to several flaws commonly found in real datasets : false negatives, false positives and parameter heterogeneity.
Parameter heterogeneity was only studied for colonization probability by assuming that a proportion of patches had a colonization probability equal to 0. Details of the protocols are provided in Supporting Information C.2.

5.2.4 Applying SBCE probability to real-world monitoring data

In order to apply SBCE probability to real-world presence/absence datasets of plant metapopulations, we used the following protocol. This protocol can be applied to any metapopulation of annual plants, as well as any plant metapopulation in which the plants are killed yearly.

- *Step 1 :* Choose which subsets of the metapopulation are expected to evolve according to the same parameter values.
- *Step 2 :* Choose whether the patches that are never occupied have a colonization probability equal to 0. If so, then exclude these patches from analysis.
- *Step 3 :* Exclude from analysis the metapopulation subsets that contain too few patches (typically less than 20).
- *Step 4 :* Perform SBCE probability estimation on each metapopulation subset.

The code provided with the article can be used to estimate the SBCE probability and to compute an associated 95% confidence interval, and works with Python 3.3 or later versions. See Section 5.5.5 for a tutorial on how to use this code.
The real dataset used in this study, *Paris 12*, consists of floristic inventories of 1324 tree bases located in Paris 12th administrative district, carried out annually between 2009 and 2018. Spontaneous flora was inventoried exhaustively over the entire period, except in 2013, when a limited number of species were tracked. The taxonomic reference is the French Flora Reference TAXREF v8.0 [Gar+14]. See Supporting Information (B.1) or [Oma+18] for information about nomenclature and the species and streets monitored.
Standing plants were removed yearly by the gardners as part of the tree bases management. We considered that the tree bases present in different streets represented distinct metapopulations. Moreover, we considered that the metapopulation parameters could be different from one species to another, and for a given species, from one street (i.e one metapopulation) to another. For almost every species, a high proportion of patches were never occupied over the study period. We interpreted this as being due to colonization heterogeneity, and considered that patches which were never colonized had a colonization probability equal to 0. Since the germination and seed bank death probabilities could not be estimated for a group of patches that were never occupied, we removed these patches from the analysis. We then retained each pair of species and street containing at least 20 patches left (the size of a street ranged from 31 to 186 tree bases). For each pair, seed bank contribution was assessed under the hypothesis that the species' patch occupancy dynamics followed the Propagule Rain Model. Parameter estimation for SB+ model was carried out using the missing data variant when the species was not inventoried in 2013 (see Section 5.5.1).
For each species, we tested the hypothesis that metapopulation parameters were different from one street to another by comparing AICs of parameter fits assuming that all 4 parameters were respectively different or identical from one street to another. We used the Holm-Bonferroni method with a 0.05 significance level to account for simultaneous testing.
For each pair of species and street, we performed a non-parametric bootstrap analysis by generating 1000 new datasets, to compute the distribution of the estimated SBCE probability and the 95% confidence interval on $\mathbb{P}_{SBCE}$ estimation.
We then tested whether the estimated SBCE probability was affected by the nature of the closest green space (see [Oma+19]) the seed dispersal mechanism, the flowering months (extracted from the database of the collaborative network of French botanists "Tela botanica" (http://www.tela-botanica.org), the releasing height of the seeds and the seed weight (mean value obtained from the LEDA database [Kle+08]). We used mixed-effect linear regression model with SBCE probability as response wariable, in which we integrated a weighting vector to take into account the degree of

confidence on the estimated SBCE probability. See Section 5.5.4 for the details of the protocol.
Besides the above regression model, we also performed repeatability analyses to provide an overview of the variation of the germination probability and SBCE probability among streets and among species. The repeatability analysis was based on 1000 parametric bootstraps as implemented in the rptR package of R [SNS17]. The statistical significance of the repeatability of each metric was tested by a likelihood ratio test comparing the model fit of a model including grouping factor (here, the species or the street) and one excluding it.

5.3 Results

5.3.1 Criterium for seed bank identification

For SB+ models, the Mean Absolute Error (MAE) on SBCE estimation is presented in Figure 5.4. It reached values below 0.05 when 200 (resp. 100) patches were monitored during 5 (resp. 10) years, and below 0.025 when 500 (resp. 800) patches were observed during 10 (resp. 5) years. For SB- models, for each monitoring duration, we could find a number of patches ensuring $p_{fms} < 0.05$. The minimal number of patches required decreased from 200 for a monitoring lasting 3 years to 30 for a monitoring lasting 20 years. Complete results can be found in Supporting Information C.1.

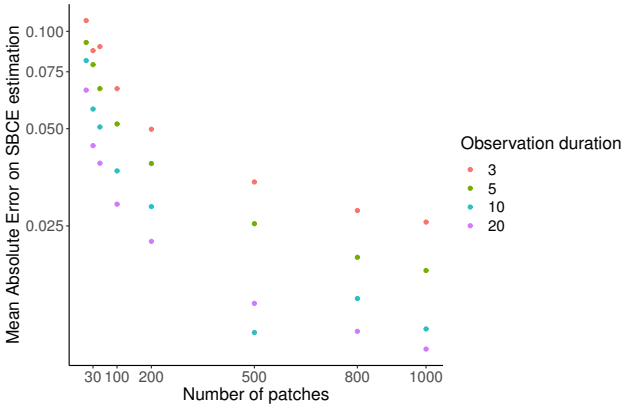


Figure 5.4: Mean Absolute Error (MAE) on SBCE probability estimation for SB+ models (logarithmic scale). 30 datasets were generated for each combination of parameter values, number of patches and years of observation as presented in Table 5.1.

5.3.2 Testing SBCE estimation robustness

Overall, the introduction of false negatives, false positives or heterogeneous colonization increased the bias on SBCE probability estimation for SB+ models, and p_{fms} for SB- models.
For SB+ models, the introduction of false negatives or false positives had contrasted effects depending on the parameter sets. For some parameter sets, the bias increased with the error rate, leading to an overestimation or underestimation (depending on the parameter set) of SBCE probability. For other parameter sets, the bias increased less with the error rate, and increasing the monitoring duration or the number of patches led to a marked decrease of the bias. Conversely, the introduction of heterogeneous colonization (i.e, when we set the colonization probability of some patches to 0) led to the SBCE probability being overestimated, even for a proportion of non-colonizable patches of 0.05, especially for long monitoring durations.

For SB- models, when considering false negatives or false positives, increasing the monitoring duration or the number of patches led to a decrease of p_{fms} (albeit less marked for a false negative rate of 0.2), and the 0.05 threshold could be met for a sufficient monitoring duration or number of patches, except for a false negative rate of 0.2 (see Table 5.2). Conversely, when considering heterogeneous colonization, increasing the number of patches had no marked effect, and increasing the monitoring duration actually made p_{fms} increase, reaching values of up to 0.5. Consequently, no combination of number of years of observation and number of patches monitored fullfilled the accuracy requirements, even for a proportion of non-colonizable patches as low as 0.05.
Complete results can be found in Supporting Information C.3.

Table 5.2: Minimal number of patches needed to satisfy $p_{fms} < 0.05$ for SB- models, after having introduced false negatives or false positives in the dataset, for different monitoring durations and rates of false negatives or false positives. The symbol X means that satisfaction criterium could not be met.

Monitoring duration	**3 years**	**5 years**	**10 years**	**20 years**
False negative rate	**Number of patches**			
0	500	200	100	30
0.05	1000	500	500	200
0.1	X	1000	500	500
0.2	X	X	X	X
False positive rate	**Number of patches**			
0	500	200	100	30
0.05	X	1000	500	100
0.1	X	X	500	500
0.2	X	X	X	800

5.3.3 Applying SBCE probability to real-world monitoring data

The analysis highlighted a high variability of SBCE probabilities, both between species and within species. A medium to strong seed bank effect was detected in at least one street for 23 species out of the 46 species considered, while a weak seed bank effect was detected in at least one street for 17 species (see Table 5.3 and Section 5.5.4). Germination probabilities were generally fairly low, most of the time below 0.5, no matter the species or the street. Estimations of SBCE and germination probabilities can be found in Supporting Information D.2.4.

The null hypothesis of identical metapopulation parameters from one street to another was rejected for all but 5 species (see Supporting Information D.2.1). Moreover, the repeatabilities of the germination probability g and SBCE probability between species for a given street were low, although significantly different from zero (repeatability $R = 0.053 \pm 0.03$, p-value $= 0.00312$ for germination and $R = 0.045 \pm 0.04$, p-value $= 0.0427$ for SBCE). In contrast, repeatabilities of these probabilities were relatively high between streets for a given species (repeatability $R = 0.56 \pm 0.08$, p-value $< 10^{-4}$ for germination and $R = 0.192 \pm 0.08$, p-value $< 10^{-5}$ for SBCE).
According to the regression model, none of our explaining variables was significantly correlated with the SBCE probability (see detailed regression results in Supporting Information D.2.3).

Table 5.3: Species for which the 95% confidence interval on the estimated SBCE probability was completely included in $[0, 0.05]$ (*weak seed bank effect*) or $[0.05, 1]$ (*medium to strong seed bank effect*).

Weak seed bank effect		**Medium to strong seed bank effect**	
Apera spica-venti	Chenopodium album	Amaranthus retroflexus	Capsella bursa-pastoris
Cirsium vulgare	Conyza	Cardamine hirsuta	Carduus pycnocephalus
Hordeum murinum	Lactuca serriola	Cerastium glomeratum	Chenopodium album
Matricaria	Plantago lanceolata	Cirsium arvense	Conyza
Plantago major	Poa annua	Elytrigia repens	Hordeum murinum
Polygonum persicaria	Senecio vulgaris	Lactuca muralis	Lactuca serriola
Sinapis arvensis	Sisymbrium irio	Matricaria	Oxalis corniculata
Sonchus	Stellaria media	Poa annua	Polygonum aviculare
Taraxacum		Sedum vulgare	Senecio vulgaris
		Sisymbrium irio	Sisymbrium officinale
		Sonchus	Stellaria media
		Taraxacum	Veronica persica

5.4 Discussion

In this paper, we propose a new metric, the seed bank characteristic event probability (SBCE) providing information on the contribution of the seed bank to the observed standing vegetation dynamics in plant metapopulations. Our results indicated that the SBCE performs well in a wide range of situations and provided evidence of a significant contribution of the seed bank to plant metapopulation dynamics in an urban environment. In biology, as in other disciplines, it is sometimes more practical and straightforward to make inferences about a process within a system by observing the effects of that process on the system rather than the process itself. This approach is central to the study of metapopulations, where a process such as dispersal is often studied indirectly by examining its consequences in terms of genetic structuring or recolonization dynamics. This idea is also at the core of the analytical framework recently developed [Fré+13; Bor+15] to study seed banks, which are very difficult to detect on a large scale, but whose consequences in terms of plant metapopulation dynamics can be crucial [Fen95].

5.4.1 Theoretical analysis

Our analysis indicates that the SBCE approach is relevant to evaluate the contribution of a potential seed bank to the dynamics of a metapopulation in which patches are independent (i.e., in which the colonization or extinction of a patch does not depend on the presence of the species in the other patches), even when the number of monitored patches and monitoring duration are limited. Indeed, for a sufficient but still achievable monitoring duration or number of patches, the SBCE probability was well estimated when seed bank existed and correctly reflected the absence of a seed bank otherwise (see Supporting Information C.1). The SBCE approach can be used on metapopulations of annual plant species, as well as metapopulations in which all plants are killed after having produced seeds.

Contrary to traditionnal methods used to assess directly the presence of a seed bank, which involve putting soil samples in germination chambers or using seed separation methods [BB14], or to the estimation method introduced in [Plu+18], the SBCE approach is not suited for concluding to the presence of a seed bank, nor for getting estimates of the seed bank parameters, namely probabilities of germination and survival of seeds. However, monitoring the species presence/absence is easier than assessing directly the presence of the seed bank by laboratory experiments. More-

over, when the seed bank only contributes marginally to the observed metapopulation dynamics, estimating accurately the seed bank parameters under [Plu+18] is data-demanding. Our study suggests that if one is interested in assessing seed bank contribution to the global observed dynamics rather than seed bank parameters or seed bank presence, then the SBCE approach requires less data than the estimation of seed bank parameters and can be used in a broader range of real-life situations.

Nevertheless, our work highlights that with strongly heterogeneous colonization (e.g., in the special case in which a fraction of patches can never be colonized), the SBCE probability cannot be well estimated, no matter the monitoring duration or the number of patches. Therefore it is crucial to identify heterogeneous colonization situations, whose impact on estimation accuracy cannot be mitigated by improving the monitoring effort. Other statistical methods can be used to identify heterogeneous colonization cases, allowing to treat the case where colonization probabilities are heterogeneous but potentially all non-zero, for instance using mixture models [Rob18]. However these methods are far more computationally intensive than the simple one we used.

On the other hand, false negatives or false positives in standing vegetation detection have a weaker impact on the estimation of the SBCE probability. Indeed, a moderate presence of false positives or false negatives can be mitigated by monitoring more patches over a longer duration when the error rate is not too high. As a result, contrary to parameters classically considered [Moi02], the SBCE probability can still be well estimated when the false positive or false negative rate is nonzero but low (below 0.05 or even 0.1), provided enough patches are monitored during a sufficient duration. For higher rates, alternative methods are needed, such as ones accounting for the presence of false negatives [Moi02] or detectability [LPR18].
Our results suggest that the method may be used on data coming from citizen science programs, for which the number of patches monitored is typically very large. Recent empirical results on citizen science programs showed that, with a standardized protocol and good training methods, error rates in datasets can remain below 0.05 [Fuc+15; Rat+16], which makes these programs compatible with our SBCE approach.

5.4.2 The Paris 12 dataset

The analysis performed on the Paris 12 dataset of plants present in tree bases in Paris showed that 22% of the pairs of species and streets analysed exhibited a medium to strong seed bank effect, suggesting that seed banks have a key influence on plant metapopulation dynamics in this type of urban environment. Estimates of both germination probability g and SBCE probability are highly consistent between streets for a given species, even though for 83% of the species present in at least one street, metapopulation parameters are significatively different between streets. Conversely, these estimates are less consistent from one species to another for a given street, suggesting that both germination and SBCE probabilities primarily depend on species rather than location.
We did not find any relationship between SBCE estimates and the species and environmental variables tested (type of green space closest to the street, flowering period, dispersal mechanism and weight of the seeds). These last two results are in line with the findings of [Oma+18], who showed that the distribution of species in Paris tree bases was not correlated to the dispersal mechanism nor to the weight of the seeds, and hypothetised that this was partly due to human activity spreading all seeds no matter the weight or the dispersal device [Suk04; VK07].
The applicability of SBCE probability on the *Paris 12* dataset has some limitations. Indeed, we assumed that tree bases' dynamics are independent of each other (i.e., no colonization from one patch to another). As a result, we did not consider the alternative hypothesis of Levins metapopulation model [Lev69] with a seed bank. In this model, colonization events do not bring seeds from a propagule rain, but from neighbouring patches. Therefore, colonization probability is not the same

for all patches, and depends on the state of the metapopulation. An estimation method for this model exists [LCP19], but uses abundance data instead of presence/absence data, and would need to be adapted to handle missing data. Moreover, the consistency of g estimates for a given species between streets suggests that the assumption of independent patches was correct.
Overall, our results show that measuring seed bank contribution to plant metapopulation dynamics is less data-demanding than assessing seed bank presence, while being robust to the presence of false negatives or false positives. Our method can be applied to a wide range of urban and non-urban metapopulations, and can be used on datasets collected using citizen science in order to increase substantially the understanding of plant metapopulation dynamics.

5.5 Supporting Information

5.5.1 Variant with missing data of the estimation method

Let N be the number of patches, and T the monitoring duration. For all $t \in [\![1, T]\!]$ and $n \in [\![1, N]\!]$, let $X_t^i \in \{0, 1\}$ denote the presence/absence of plants in patch i during the generation t, and let $Z_t^i \in \{(0, 0), (1, 0), (1, 1)\}$ denote the presence/absence of seeds in patch i at the beginning of generation t, and of plants in patch i at generation t.
For all $i \in [\![1, N]\!]$, $(Z_t^i)_{1\leq T}$ is a Markov chain. Moreover, if the metapopulation parameters are $\theta = (p_0, g, c, d)$, for all $1 \leq i \leq N$,

$$\begin{aligned}\mathbb{P}(Z_1^i = (0, 0)) &= 1 - p_0\\ \mathbb{P}(Z_1^i = (1, 0)) &= p_0 \times (1 - g)\\ \mathbb{P}(Z_1^i = (1, 1)) &= p_0 \times g\end{aligned}$$

and the transition matrix of $(Z_t^i)_{1\leq t\leq T}$, denoted Q^θ, is

	$(0,0)$	$(1,0)$	$(1,1)$
$(0,0)$	$1-c$	$c(1-g)$	cg
$(1,0)$	$d(1-c)$	$(1-d(1-c))(1-g)$	$(1-d(1-c))g$
$(1,1)$	0	$1-g$	g

or equivalently, letting $c_2 = 1 - d(1 - c)$,

	$(0,0)$	$(1,0)$	$(1,1)$
$(0,0)$	$1-c$	$c(1-g)$	cg
$(1,0)$	$1-c_2$	$c_2(1-g)$	c_2g
$(1,1)$	0	$1-g$	g

Let $A \in \{E \in \mathcal{P}([\![1, T]\!]), 1 \in E\}$ be the years during which the presence/absence of plants in patches is observed. Let $X_{obs}^i = (X_t^i)_{t\in A}$ the *observed* presence/absence of plants in patch i over the period A.

We are looking for the parameter tuple $\theta = (p_0, g, c, d)$ maximizing observation likelihood :

$$\begin{aligned}\theta^{max} &= \underset{\theta}{argmax}\ l((X_{obs}^i)_{1\leq i\leq N}|\theta)\\ &= \underset{\theta}{argmax}\ \left(\mathbb{E}_\theta\left[l\left((X_{obs}^i)_{1\leq i\leq N}, (Z_t^i)_{\substack{1\leq i\leq N,\\ 1\leq t\leq T}}|\theta\right)|(X_{obs}^i)_{1\leq i\leq N}\right]\right.\\ &\qquad\left.-\mathbb{E}_\theta\left[l\left((Z_t^i)_{\substack{1\leq i\leq N,\\ 1\leq t\leq T}}|(X_{obs}^i)_{1\leq i\leq N}, \theta\right)|(X_{obs}^i)_{1\leq i\leq N}\right]\right)\end{aligned}$$

This quantity cannot be computed directly, but the Expectation Maximization (EM) algorithm (Dempster 1977) can be used to approach θ_{max}. This algorithm is iterative, giving at each time step a new value for θ which increases the likelihood. During the k-th step, if θ^k is the estimate for θ_{max} at the beginning of the step,

- during the expectation (E) step, the quantity

$$\mathbb{E}_{\theta^k}\left[l\left((X^i_{obs})_{1\leq i\leq N},(Z^i_t)_{\substack{1\leq i\leq N,\\1\leq t\leq T}}|\theta\right)|(X^i_{obs})_{1\leq i\leq N}\right]$$

  is computed for an unknown variable θ.

- during the maximization step (M) step, we compute :

$$\theta^{k+1}=\underset{\theta}{argmax}\ \mathbb{E}_{\theta^k}\left[l\left((X^i_{obs})_{1\leq i\leq N},(Z^i_t)_{\substack{1\leq i\leq N,\\1\leq t\leq T}}|\theta\right)|(X^i_{obs})_{1\leq i\leq N}\right]$$

We start by the expectation step. By independence of patches, if $\theta=(p_0,g,c,d)$,

$$\begin{aligned}
&\mathbb{E}_{\theta^k}\left[l\left((X^i_{obs})_{1\leq i\leq N},(Z^i_t)_{\substack{1\leq i\leq N,\\1\leq t\leq T}}|\theta\right)|(X^i_{obs})_{1\leq i\leq N}\right]\\
=&\sum_{i=1}^{N}\mathbb{E}_{\theta^k}\left[l(Z^i_1|\theta)|X^i_{obs})\right]+\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathbb{E}_{\theta^k}\left[l(Z^i_{t+1}|Z^i_t,\theta)|X^i_{obs}\right]\\
=&\sum_{i=1}^{N}\big(\log(1-p_0)\times\mathbb{P}_{\theta^k}(Z^i_1=(0,0)|X^i_{obs})+\log(p_0(1-g))\times\mathbb{P}_{\theta^k}(Z^i_1=(1,0)|X^i_{obs})\\
&\qquad+\log(p_0\times g)\times\mathbb{P}_{\theta^k}(Z^i_1=(1,1)|X^i_{obs})\big)\\
&+\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{z,z'=(0,0),(1,0),(1,1)}\mathbb{P}_{\theta^k}(Z^i_t=z,Z^i_{t+1}=z'|X^i)\times\log(Q^{\theta^k}(z,z'))\\
=&\log(1-p_0)\times\left(\sum_{i=1}^{N}\mathbb{P}_{\theta^k}(Z^i_1=(0,0)|X^i_{obs})\right)\\
&+\log(p_0)\times\left(\sum_{i=1}^{N}\mathbb{P}_{\theta^k}(Z^i_1=(1,0)|X^i_{obs})+\mathbb{P}_{\theta^k}(Z^i_1=(1,1)|X^i_{obs})\right)\\
&+\log(1-g)\times\Bigg(\sum_{i=1}^{N}\Bigg[\mathbb{P}_{\theta^k}(Z^i_1=(1,0)|X^i_{obs})+\sum_{t=1}^{T-1}\big(\mathbb{P}_{\theta^k}(Z^i_t=(0,0),Z^i_{t+1}=(1,0)|X^i_{obs})\\
&\qquad+\mathbb{P}_{\theta^k}(Z^i_t=(1,0),Z^i_{t+1}=(1,0)|X^i_{obs})\\
&\qquad+\mathbb{P}_{\theta^k}(Z^i_t=(1,1),Z^i_{t+1}=(1,0)|X^i_{obs})\big)\Bigg]\Bigg)\\
&+\log(g)\times\Bigg(\sum_{i=1}^{N}\Bigg[\mathbb{P}_{\theta^k}(Z^i_1=(1,1)|X^i_{obs})+\sum_{t=1}^{T-1}\big(\mathbb{P}_{\theta^k}(Z^i_t=(0,0),Z^i_{t+1}=(1,1)|X^i_{obs})\\
&\qquad+\mathbb{P}_{\theta^k}(Z^i_t=(1,0),Z^i_{t+1}=(1,1)|X^i_{obs})\\
&\qquad+\mathbb{P}_{\theta^k}(Z^i_t=(1,1),Z^i_{t+1}=(1,1)|X^i_{obs})\big)\Bigg]\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&+ \log(1-c) \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathbb{P}_{\theta^k}(Z_t^i = (0,0), Z_{t+1}^i = (0,0) | X_{obs}^i) \right) \\
&+ \log(c) \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathbb{P}_{\theta^k}(Z_t^i = (0,0), Z_{t+1}^i = (1,0) | X_{obs}^i) + \mathbb{P}_{\theta^k}(Z_t^i = (0,0), Z_{t+1}^i = (1,1) | X_{obs}^i) \right) \\
&+ \log(1-c') \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathbb{P}_{\theta^k}(Z_t^i = (1,0), Z_{t+1}^i = (0,0) | X_{obs}^i) \right) \\
&+ \log(c') \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathbb{P}_{\theta^k}(Z_t^i = (1,0), Z_{t+1}^i = (1,0) | X_{obs}^i) + \mathbb{P}_{\theta^k}(Z_t^i = (1,0), Z_{t+1}^i = (1,1) | X_{obs}^i) \right),
\end{aligned}
$$

which is of the form

$$a_{p_0} \log(p_0) + b_{p_0} \log(1-p_0) + a_c \log(c) + b_c \log(1-c) + a_g \log(g) + b_g \log(1-g) + a_{c'} \log(c') + b_{c'} \log(1-c')$$

In other words, if we can compute the probabilities $\mathbb{P}_{\theta^k}(Z_t^i = z | X_{obs}^i)$ and $\mathbb{P}_{\theta^k}(Z_t^i = z, Z_{t+1}^i = z' | X_{obs}^i)$ for all $1 \leq i \leq N$, $1 \leq t \leq T-1$ and $z, z' \in \{(0,0), (1,0), (1,1)\}$, then we can maximize

$$\mathbb{E}_{\theta^k} \left[l \left((X_{obs}^i)_{1 \leq i \leq N}, (Z_t^i)_{\substack{1 \leq i \leq N, \\ 1 \leq t \leq T}} | \theta \right) | (X_{obs}^i)_{1 \leq i \leq N} \right]$$

and compute θ^{k+1}. Therefore, let $1 \leq i \leq N$, $1 \leq t \leq T-1$ and $z, z' \in \{(0,0), (1,0), (1,1)\}$.

$$
\begin{aligned}
\mathbb{P}_{\theta^k}(Z_t^i = z | X_{obs}^i) &= \sum_{X_{mis}^i \in \{0,1\}^{[\![1,T]\!] \setminus A}} \mathbb{P}_{\theta^k}(Z_t^i = z, X_{mis}^i | X_{obs}^i) \\
&= \sum_{X_{mis}^i \in \{0,1\}^{[\![1,T]\!] \setminus A}} \mathbb{P}_{\theta^k}(X_{mis}^i | X_{obs}^i) \times \mathbb{P}_{\theta^k}(Z_t^i = z | X_{obs}^i \cup X_{mis}^i) \\
&= \sum_{X_{mis}^i \in \{0,1\}^{[\![1,T]\!] \setminus A}} \frac{\mathbb{P}_{\theta_k}(X_{obs}^i \cup X_{mis}^i)}{\mathbb{P}_{\theta_k}(X_{obs}^i)} \times \frac{\mathbb{P}_{\theta_k}(Z_t^i = x, X_{obs}^i \cup X_{mis}^i)}{\mathbb{P}_{\theta^k}(X_{obs}^i \cup X_{mis}^i)} \\
&= \sum_{X_{mis}^i \in \{0,1\}^{[\![1,T]\!] \setminus A}} \frac{\mathbb{P}_{\theta^k}(Z_t^i = z, X_{obs}^i \cup X_{mis}^i)}{\mathbb{P}_{\theta^k}(X_{obs}^i)}
\end{aligned}
$$

and similarly

$$\mathbb{P}_{\theta^k}(Z_t^i = z, Z_{t+1}^i = z' | X_{obs}^i) = \sum_{X_{mis}^i \in \{0,1\}^{[\![1,T]\!] \setminus A}} \frac{\mathbb{P}_{\theta^k}(Z_t^i = z, Z_{t+1}^i = z', X_{obs}^i \cup X_{mis}^i)}{\mathbb{P}_{\theta^k}(X_{obs}^i)}$$

We can then compute the probabilities $\mathbb{P}_{\theta^k}(Z_t^i = z, X_{obs}^i \cup X_{mis}^i)$, $\mathbb{P}_{\theta^k}(Z_t^i = z, Z_{t+1}^i = z', X_{obs}^i \cup X_{mis}^i)$ and $\mathbb{P}_{\theta^k}(X_{obs}^i)$ using the forward-backward algorithm (Rabiner 1989).

5.5.2 Computation of the *seed bank proportional occupancy gain*

For a metapopulation of parameters g, c, d and p_0 monitored during t years, the *seed bank proportional occupancy gain* is the proportion of patches that are occupied by plants, but would not if there was no seed bank (i.e, if $d = 1$). This includes patches occupied by plants all coming from seeds which spent at least one generation dormant (in other words, plants germinating as part of a Seed Bank Characteristic Event), but also the ones containing the descendants of such plants. Therefore the *seed bank proportional occupancy gain* is different from the SBCE probability, even though we

expect them to be close.
When $t \to +\infty$, then the *seed bank proportional occupancy gain* converges to a quantity which can be computed using the stationary distributions associated to the PRMs of respective parameters (g, c, d, p_0) and (g, c, 1, p_0) (see B.1). However, for finite monitoring durations, we have to approximate it by doing simulations (see B.2).

Limit of the *seed bank proportional occupancy gain* when $t \to +\infty$
Let ($z_{0,0}$, $z_{1,0}$, $z_{1,1}$) be the stationary distribution associated to the Markov chain of transition matrix

	$(0,0)$	$(1,0)$	$(1,1)$
$(0,0)$	$1-c$	$c(1-g)$	cg
$(1,0)$	$d(1-c)$	$(1-d(1-c))(1-g)$	$(1-d(1-c))g$
$(1,1)$	0	$1-g$	g

In other words, ($z_{0,0}$, $z_{1,0}$, $z_{1,1}$) satisfies :

(1) $z_{0,0} + z_{1,0} + z_{1,1} = 1$
(2) $z_{0,0} \times (1-c) + z_{1,0} \times d(1-c) = z_{0,0}$
(3) $z_{0,0} \times c(1-g) + z_{1,0} \times (1-d(1-c))(1-g) + z_{1,1} \times (1-g) = z_{1,0}$
(4) $z_{0,0} \times cg + z_{1,0} \times (1-d(1-c))g + z_{1,1} \times g = z_{1,1}$

From (2), we deduce :

$$z_{0,0} = z_{1,0} \times \frac{d(1-c)}{c}$$

Combining it with (4), we get

$$g \times d(1-c) \times z_{1,0} + z_{1,0} \times (1-d(1-c))g = z_{1,1} \times (1-g)$$

i.e.,

$$z_{1,0} = z_{1,1} \times \frac{1-g}{g}$$

And combining these two equations with (1), we get :

$$\begin{aligned} z_{1,1} \times \frac{1-g}{g} \times \frac{d(1-c)}{c} + z_{1,1} \times \frac{1-g}{g} + z_{1,1} &= 1 \\ \Leftrightarrow z_{1,1} \times \frac{(1-g) \times d(1-c) + c(1-g) + cg}{cg} &= 1 \\ \Leftrightarrow z_{1,1} &= \frac{cg}{c + d(1-g)(1-c)}. \end{aligned}$$

Therefore, the limit of the *seed bank proportional occupancy gain* when $t \to +\infty$ is given by

$$\begin{aligned} \frac{cg}{c + d(1-c)(1-g)} - \frac{cg}{c + (1-g)(1-c)} &= \frac{cg \times (1-g)(1-c) - cg \times d(1-g)(1-c)}{(c + d(1-g)(1-c)) \times (c + (1-c)(1-g))} \\ &= \frac{cg(1-g)(1-c)(1-d)}{(c + d(1-g)(1-c)) \times (c + (1-g)(1-c))} \end{aligned}$$

Approximation of the *seed bank proportional occupancy gain* for finite monitoring durations

In order to approximate the *seed bank proportional occupancy gain*, we considered metapopulations of 10000 patches. For a given monitoring duration t and a given initial proportion p_0, the parameters g, c and d spanned values between 0.2 and 0.9 (resp. 1, 1) with a 0.05 step, yielding 4335 different parameter sets.
For a given parameter set (g, c, d, p_0), we computed the approximate *seed bank proportional occupancy gain* by considering 30 metapopulations of 10000 patches during t consecutive years. We coupled to each original metapopulation another one the following way :

1. We started by considering 10000 patches, each being an exact copy of a patch in the original metapopulation.

2. At each generation, we updated each metapopulation the following way. For each metapopulation and each patch, if it contained seeds *and* if a germination event occured in the same patch in the original metapopulation, then the seeds germinated, giving plants which produced new seeds and died. If conversely no germination event occured in the original metapopulation, then the seeds died.

3. Moreover, for each metapopulation and each patch, if a colonization event occured in the original metapopulation, then another one occured in the patch considered in the coupled metapopulation.

With this coupling, we obtain 30 independant metapopulations, each following a SB- models of germination probability g and colonization probability c. Moreover, in each metapopulation, if a patch is occupied by plants, then the corresponding patch in the original metapopulation contains plants as well. Therefore, all patches which are occupied in the original metapopulation but not in the coupled metapopulation are patches which are empty if the seed bank is not accounted for. This allows us to compute an approximation of the *seed bank proportional occupancy gain*, as the mean number of different patches between each of the 30 metapopulations and the corresponding coupled metapopulations.

Comparison of the SBCE probability and the *seed bank proportional occupancy gain*

The following figures compare the SBCE probability and the *seed bank proportional occupancy gain*, respectively when the monitoring duration $t \to +\infty$ (Figure 5.5) and for a 10 years monitoring duration (Figure 5.6). For the finite monitoring duration case, only the cases $p_0 = 0$ and $p_0 = 1$ were considered. See Figure 5.3 in the main article for a comparison in the case $t = 50$ and $p_0 = 1$.

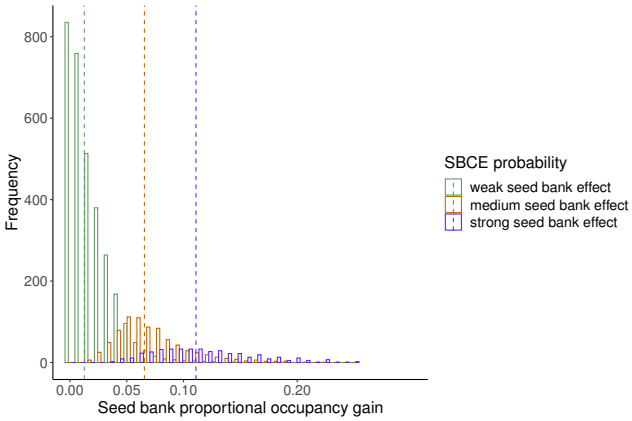
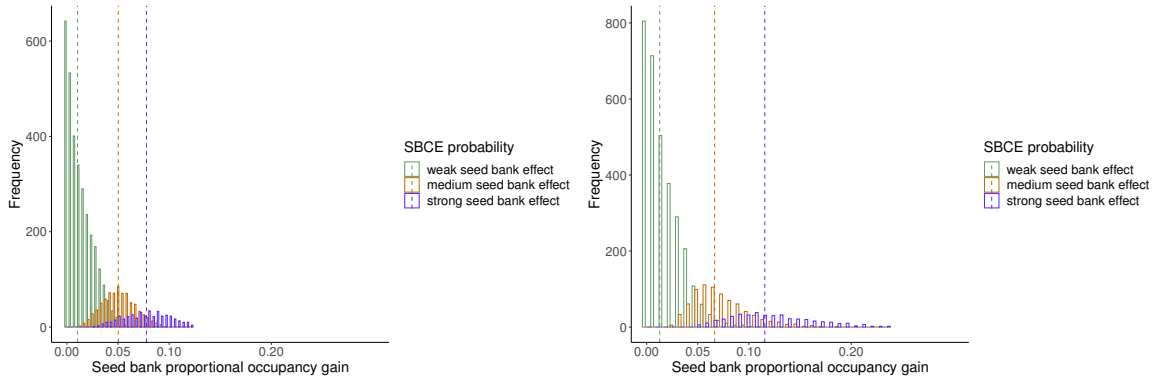


Figure 5.5: Relationship between the SBCE probability and the *seed bank proportional occupancy gain* in the limit $t \to +\infty$. We considered 4335 distinct parameter sets, presented in Section 5.5.2. For each parameter set, values of the seed bank proportional occupancy gain and the SBCE probability were computed. Parameter sets were then classified in three groups according to their SBCE probability value (weak, medium or strong seed bank effect). Dotted vertical lines indicate the median seed bank proportional occupancy gain for each group.



(a) Initial proportion of seeds $p_0 = 0$ (b) Initial proportion of seeds $p_0 = 1$

Figure 5.6: Relationship between the SBCE probability and the *seed bank proportional occupancy gain*. We modelled 4335 distinct parameter sets of large metapopulations over $t = 10$ consecutive years. For each parameter set, values of the seed bank proportional occupancy gain and the SBCE probability were computed. Parameter sets were then classified in three groups according to their SBCE probability value (weak, medium or strong seed bank effect). Dotted vertical lines indicate the median seed bank proportional occupancy gain for each group. The description of parameter sets and the method used to approximate the seed bank proportional occupancy gain are presented in Section 5.5.2.

5.5.3 Analysis of the estimator performances

Results of the theoretical study

Table 5.4: Theoretical values taken by the SBCE probability for SB+ parameter sets used in the study.

Parameters			
c	g	d	$\mathbb{P}_{SBCE}$
0.3	0.3	0.2	0.19
0.3	0.3	0.6	0.073
0.3	0.7	0.2	0.14
0.3	0.7	0.6	0.064
0.7	0.3	0.2	0.061
0.7	0.3	0.6	0.028
0.7	0.7	0.2	0.054
0.7	0.7	0.6	0.026

Table 5.5: Proportion of datasets generated with a SB- model for which the estimated SBCE probability was above 0.05, depending on the number of patches and the monitoring duration.

Patches	3 years	5 years	10 years	20 years
10	0.158	0.099	0.125	0.079
30	0.146	0.107	0.071	**0.037**
50	0.137	0.083	0.054	**0.033**
100	0.079	**0.025**	**0.020**	**0.012**
200	**0.042**	**0.037**	**0.004**	**0**
500	**0.021**	**0.012**	**0**	**0**
800	**0.004**	**0**	**0**	**0**
1000	**0**	**0**	**0**	**0**

Protocols used for analysing the effect of corrupted datasets

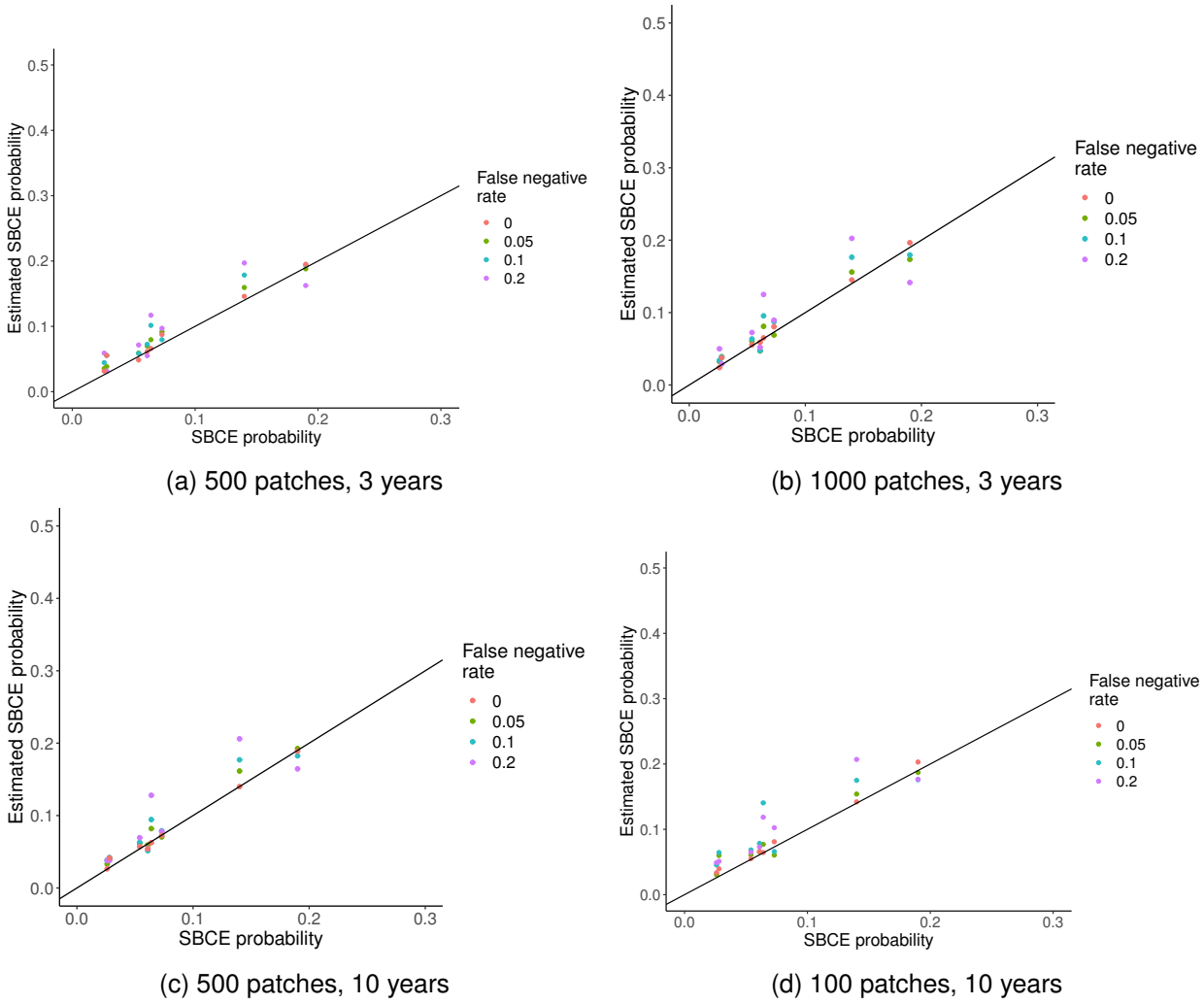
In the following, the expression *corrupted dataset* will be used for a simulated dataset into which false negatives, false positives or parameter heterogeneity have been introduced.

False negatives were defined as a species presence being wrongly recorded as an absence. The influence of false negatives was tested introducing false negatives in simulated datasets, independently over patches and time, with the same *false negative rate*, which spanned $\{0.05, 0.1, 0.2\}$. For each combination of parameters, number of patches and monitoring durations listed in Table 5.1, 30 datasets were generated. Parameter estimation with no constraint on d was performed, and SBCE probabilities were computed. Estimation accuracy was assessed computing the bias on SBCE probability estimation for each combination of germination, colonization and seed bank death probabilities. Likewise, false positives were defined as a species being wrongly recorded as present when it was not. Their influence was studied introducing false positives in simulated datasets, independently over patches and time, with the same *false positive rate*, which spanned $\{0.05, 0.1, 0.2\}$. The test protocol was the same as for false negatives.

Parameter heterogeneity influence was only studied for the colonization probability. Indeed seed bank extinction probability is already known to be poorly estimated even with simulated datasets (see (Pluntz and al., 2018)), and we expected heterogeneity in germination probability, such as lower germination probabilities during one year, to have effects close to the ones of false negatives introduction. Therefore we focused on the effect of some patches having a colonization probability equal to 0. The floristic inventories of tree bases dataset was indeed characterized by, for each species, a high proportion of patches in which it was never present.

For each combination of parameters set, number of patches and monitoring duration listed in Table 5.1, 30 datasets were generated in which some patches had been set as non colonizable, in a proportion spanning $\{0.05, 0.1, 0.2\}$. Parameter estimation with no constraint on d was performed. SBCE probabilities were computed. The precision of the estimations was assessed in the same way as for datasets with false negatives or false positives.

Tests on corrupted datasets



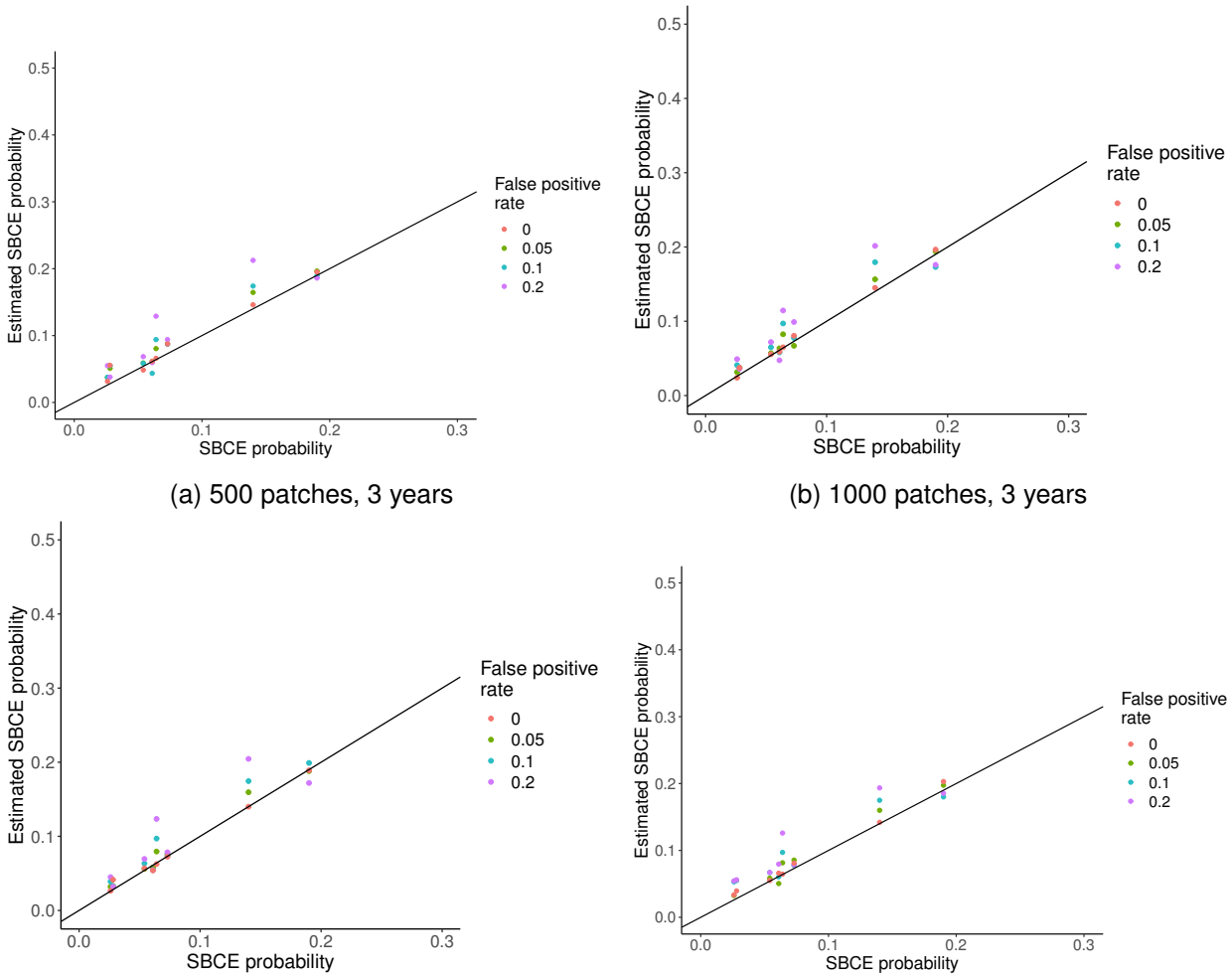
(a) 500 patches, 3 years

(b) 1000 patches, 3 years

(c) 500 patches, 10 years

(d) 100 patches, 10 years

Figure 5.7: Effect of introducing false negatives on the bias on SBCE probability estimation, for datasets generated with a SB+ model ($d < 1$). The horizontal axis corresponds to the actual SBCE probability, and the vertical axis to the mean estimated SBCE probability. The diagonal line is the bisector. For each patch and at each timestep, the probability of categorizing the patch as empty when it was occupied (*false negative rate*) ranged from 0 to 0.2.



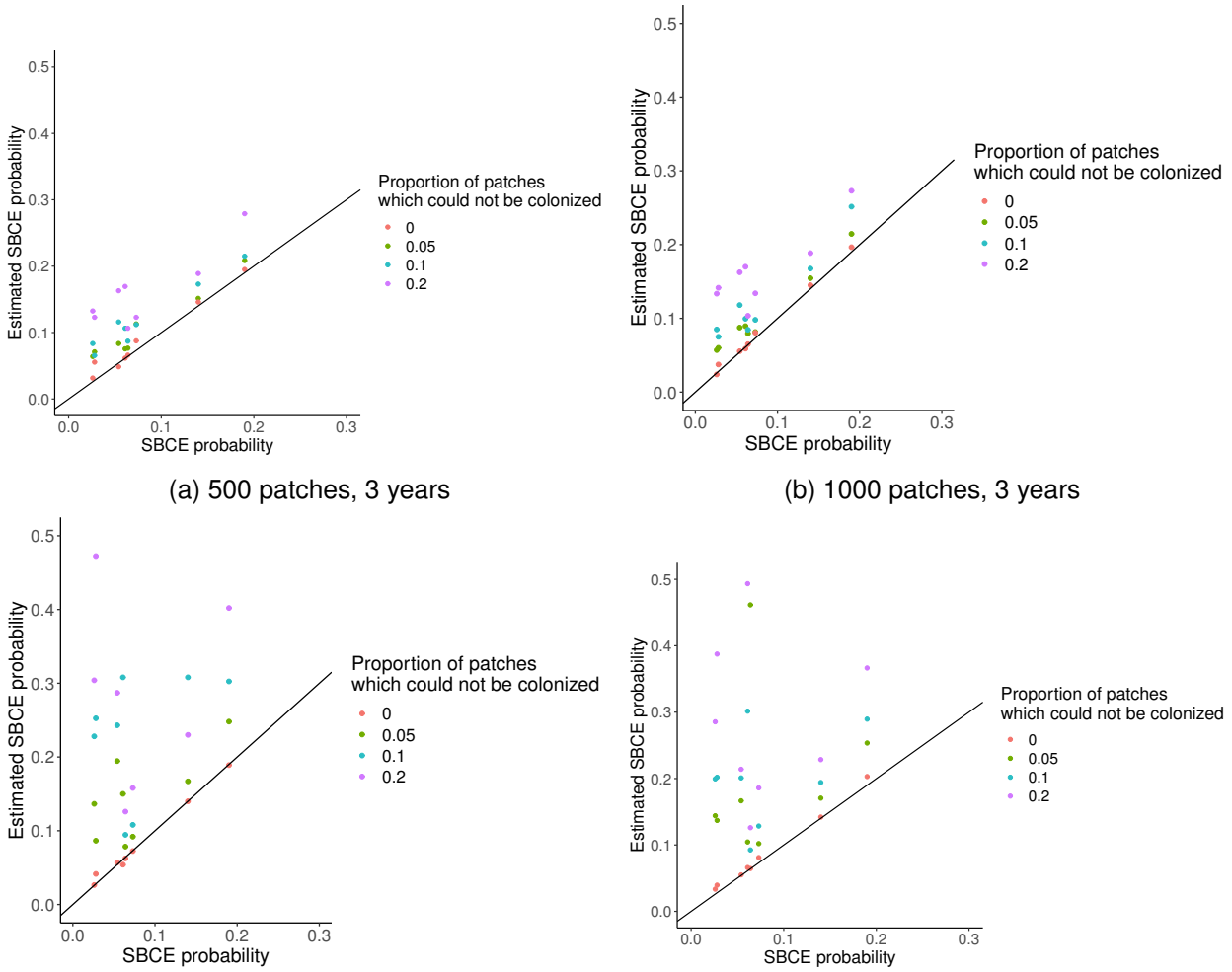
(a) 500 patches, 3 years

(b) 1000 patches, 3 years

(c) 500 patches, 10 years

(d) 100 patches, 10 years

Figure 5.8: Effect of introducing false positives on the bias on SBCE probability estimation, for datasets generated with a SB+ model ($d < 1$). The horizontal axis corresponds to the actual SBCE probability, and the vertical axis to the mean estimated SBCE probability. The diagonal line is the bisector. For each patch and at each timestep, the probability of categorizing the patch as occupied when it was empty (*false positive rate*) ranged from 0 to 0.2.



(a) 500 patches, 3 years

(b) 1000 patches, 3 years

(c) 500 patches, 10 years

(d) 100 patches, 10 years

Figure 5.9: Effect of the presence of non-colonizeable patches on the bias on SBCE probability estimation, for datasets generated following a SB + model ($d < 1$). The horizontal axis corresponds to the actual SBCE probability, and the vertical axis to the mean estimated SBCE probability. The diagonal line is the bisector. Each patch was deemed as non-colonizeable randomly and independently of other patches with a probability ranging from 0 to 0.2 (proportion of non-colonizeable patches).

False negative probability	20 years
0	***0.037***
0.05	*0.179*
0.1	*0.241*
0.2	*0.433*

(a) Dataset of 30 patches.

False negative probability	10 years	20 years
0	***0.020***	***0.012***
0.05	*0.129*	*0.100*
0.1	*0.054*	*0.121*
0.2	*0.362*	*0.304*

(b) Dataset of 100 patches.

False negative probability	5 years	10 years	20 years
0	***0.037***	***0.004***	***0***
0.05	*0.125*	*0.062*	***0.050***
0.1	*0.216*	*0.129*	*0.067*
0.2	*0.362*	*0.291*	*0.254*

(c) Dataset of 200 patches.

False negative probability	3 years	5 years	10 years	20 years
0	***0.021***	***0.012***	***0***	***0***
0.05	*0.104*	***0.042***	***0.008***	***0.025***
0.1	*0.196*	*0.100*	***0.045***	***0.004***
0.2	*0.312*	*0.283*	*0.241*	*0.271*

(d) Dataset of 500 patches.

False negative probability	3 years	5 years	10 years	20 years
0	***0.004***	***0***	***0***	***0***
0.05	*0.054*	***0.029***	***0.008***	***0.004***
0.1	*0.104*	*0.054*	***0.046***	***0.008***
0.2	*0.316*	*0.271*	*0.274*	*0.271*

(e) Dataset of 800 patches.

False negative probability	3 years	5 years	10 years	20 years
0	***0***	***0***	***0***	***0***
0.05	***0.037***	***0.025***	***0.012***	***0.004***
0.1	*0.104*	***0.046***	***0.017***	***0.020***
0.2	*0.292*	*0.254*	*0.258*	*0.266*

(f) Dataset of 1000 patches.

Table 5.6: Proportion of datasets generated with a SB- model for which the estimated SBCE probability was above 0.05, for different false negative probabilities. For each patch and at each timestep, the probability of categorizing the patch as empty when it was occupied (*false negative rate*) ranged from 0 to 0.2.

False positive probability	20 years
0	***0.037***
0.05	*0.133*
0.1	*0.225*
0.2	*0.212*

(a) Dataset of 30 patches.

False positive probability	10 years	20 years
0	***0.020***	***0.012***
0.05	*0.146*	***0.05***
0.1	*0.121*	*0.083*
0.2	*0.167*	*0.150*

(b) Dataset of 100 patches.

False positive probability	5 years	10 years	20 years
0	***0.037***	***0.004***	***0***
0.05	*0.113*	*0.087*	*0.067*
0.1	*0.129*	*0.108*	*0.054*
0.2	*0.212*	*0.125*	*0.087*

(c) Dataset of 200 patches.

False positive probability	3 years	5 years	10 years	20 years
0	***0.021***	***0.012***	***0***	***0***
0.05	*0.120*	***0.049***	***0.033***	***0.021***
0.1	*0.112*	*0.100*	***0.037***	***0.012***
0.2	*0.162*	*0.154*	*0.092*	*0.058*

(d) Dataset of 500 patches.

False positive probability	3 years	5 years	10 years	20 years
0	***0.004***	***0***	***0***	***0***
0.05	***0.058***	*0.054*	***0.033***	***0.017***
0.1	*0.095*	*0.070*	***0.025***	***0.008***
0.2	*0.129*	*0.096*	*0.054*	***0.025***

(e) Dataset of 800 patches.

False positive probability	3 years	5 years	10 years	20 years
0	***0***	***0***	***0***	***0***
0.05	*0.079*	***0.025***	***0.012***	***0.016***
0.1	*0.054*	*0.062*	***0.029***	***0.008***
0.2	*0.104*	*0.100*	*0.054*	***0.033***

(f) Dataset of 1000 patches.

Table 5.7: Proportion of datasets generated with a SB- model for which the estimated SBCE probability was above 0.05, for different false positive probabilities. For each patch and at each timestep, the probability of categorizing the patch as occupied when it was empty (*false positive rate*) ranged from 0 to 0.2.

Non-colonizeable patches	20 years
0	***0.037***
0.05	*0.267*
0.1	*0.392*
0.2	*0.495*

(a) Dataset of 30 patches.

Non-colonizeable patches	10 years	20 years
0	***0.020***	***0.012***
0.05	*0.333*	*0.275*
0.1	*0.404*	*0.442*
0.2	*0.537*	*0.529*

(b) Dataset of 100 patches.

Non-colonizeable patches	5 years	10 years	20 years
0	***0.037***	***0.004***	***0***
0.05	*0.233*	*0.233*	*0.317*
0.1	*0.358*	*0.421*	*0.446*
0.2	*0.487*	*0.521*	*0.517*

(c) Dataset of 200 patches.

Non-colonizeable patches	3 years	5 years	10 years	20 years
0	***0.021***	***0.012***	***0***	***0***
0.05	*0.158*	*0.154*	*0.200*	*0.274*
0.1	*0.250*	*0.288*	*0.395*	*0.470*
0.2	*0.396*	*0.471*	*0.533*	*0.509*

(d) Dataset of 500 patches.

Non-colonizeable patches	3 years	5 years	10 years	20 years
0	***0.004***	***0***	***0***	***0***
0.05	*0.088*	*0.158*	*0.208*	*0.312*
0.1	*0.270*	*0.279*	*0.408*	*0.467*
0.2	*0.375*	*0.458*	*0.512*	*0.500*

(e) Dataset of 800 patches.

Non-colonizeable patches	3 years	5 years	10 years	20 years
0	***0***	***0***	***0***	***0***
0.05	*0.133*	*0.125*	*0.199*	*0.313*
0.1	*0.245*	*0.287*	*0.408*	*0.458*
0.2	*0.379*	*0.458*	*0.512*	*0.500*

(f) Dataset of 1000 patches.

Table 5.8: Proportion of datasets generated with a SB- model for which the estimated SBCE probability was above 0.05, for different proportions of non colonizable patches. Each patch was randomly and independently from other patches deemed as non-colonizeable with a probability ranging from 0 to 0.2.

5.5.4 The Paris 12 dataset

Presentation of the dataset

Parisian streets retained for analysis are presented in Table 5.9. Synonymous species names within the dataset were identified, and merged for frequent enough species. Unidentified species were removed from the database. In order to reduce the rate of false negatives and false positives, we merged taxa that were difficult to distinguish from one another and for which misidentification had seemed to happen. The list of merged taxa, and the name chosen for the new group, are indicated in Table 5.10.

Table 5.9: Streets retained for analyses

Street names	Abbreviation	Closest green space	Number of tree bases
Rue Baron le Roy	BARO	Railways (R)	62
Place du Bataillon du Pacifique	BATA	Railways (R)	31
Boulevard de Bercy (part 1)	BERC	Railways (R)	126
Boulevard de Bercy (part 2)	BERY	Seine river (S)	99
Rue de Charenton	CHAR	Railways (R)	144
Rue Daumesnil	DAUM	René Dumont footpath (F)	186
Rue Joseph Kessel	KESS	Bercy park (B)	69
Place Lachambeaudie	LACH	Railways (R)	31
Rue Montgallet	MONT	René Dumont footpath (F)	52
Rue Pommard	POMM	Bercy park (B)	39
Quai de la Rapée	RAPE	Seine river (S)	97
Rue de Bercy	RBER	Railways (R)	136
Rue de Reuilly	REUI	René Dumont footpath (F)	145
Rue Taine	TAIN	Railways (R)	62

Table 5.10: Species retained for analyses

Species name	Number of streets	Species name	Number of streets
Apera spica-venti	1	*Arabidopsis thaliana*	1
Arenaria serpyllifolia	1	*Bellis perennis*	1
Capsella bursa-pastoris	9	*Cardamine hirsuta*	1
Carduus pycnocephalus	1	*Cerastium glomeratum*	2
Chenopodium album	10	*Cirsium arvense*	4
Cirsium vulgare	1	*Convolvulus arvensis*	3
Elytrigia repens	1	*Fallopia convolvulus*	2
Galium aparine	2	*Geranium rotundifolium*	2
Lactura muralis	2	*Lactuca serriola*	7
Linaria vulgaris	1	*Malva neglecta*	1
Oxalis corniculata	1	*Papaver rhoeas*	1
Parietaria judaica	2	*Plantago lanceolata*	4
Plantago major	9	*Poa annua*	14
Rumex obtusifolium	3	*Sedum album*	1
Sedum vulgare	2	*Senecio inaequidens*	8
Senecio vulgaris	11	*Sinapis arvensis*	1
Sisymbrium irio	9	*Sisymbrium officinale*	4
Stellaria media	13	*Torilis japonica*	1

Group name	Number of streets	Species merged in the group
Amaranthus retroflexus	2	*Amaranthus retroflexus*, *Amaranthus deflexus*
Conyza	14	*Conyza*, *Conyza canadensis*, *Conyza sumatrensis*, *Erigeron canadensis*
Epilobium tetragonum	1	*Epilobium*, *Epilobium tetragonum*
Hordeum murinum	10	*Hordeum murinum*, *Hordeum murale*, *Hordeum leporinum*
Matricaria	9	*Matricaria*, *Matricaria discoidea*, *Matricaria perforata*, *Matricaria recutita*
Polygonum aviculare	9	*Polygonum arenastrum*, *Polygonum aviculare*
Polygonum persicaria	2	*Persicaria maculosa*, *Polygonum persicaria*
Sonchus	14	*Sonchus*, *Sonchus asper*, *Sonchus oleraceus*, *Sonchus arvensis*
Taraxacum	14	*Taraxacum*, *Taraxacum campylodes*, *Taraxacum castaneum*, *Taraxacum officinale*
Veronica persica	7	*Veronica*, *Veronica arvensis*, *Veronica persica*, *Veronica bungabecca* *Veronica chamaedrys*, *Veronica cymbalaria*, *Veronica hederifolia*, *Veronica pettersii* *Veronica petiolata*, *Veronica serpyfollia*

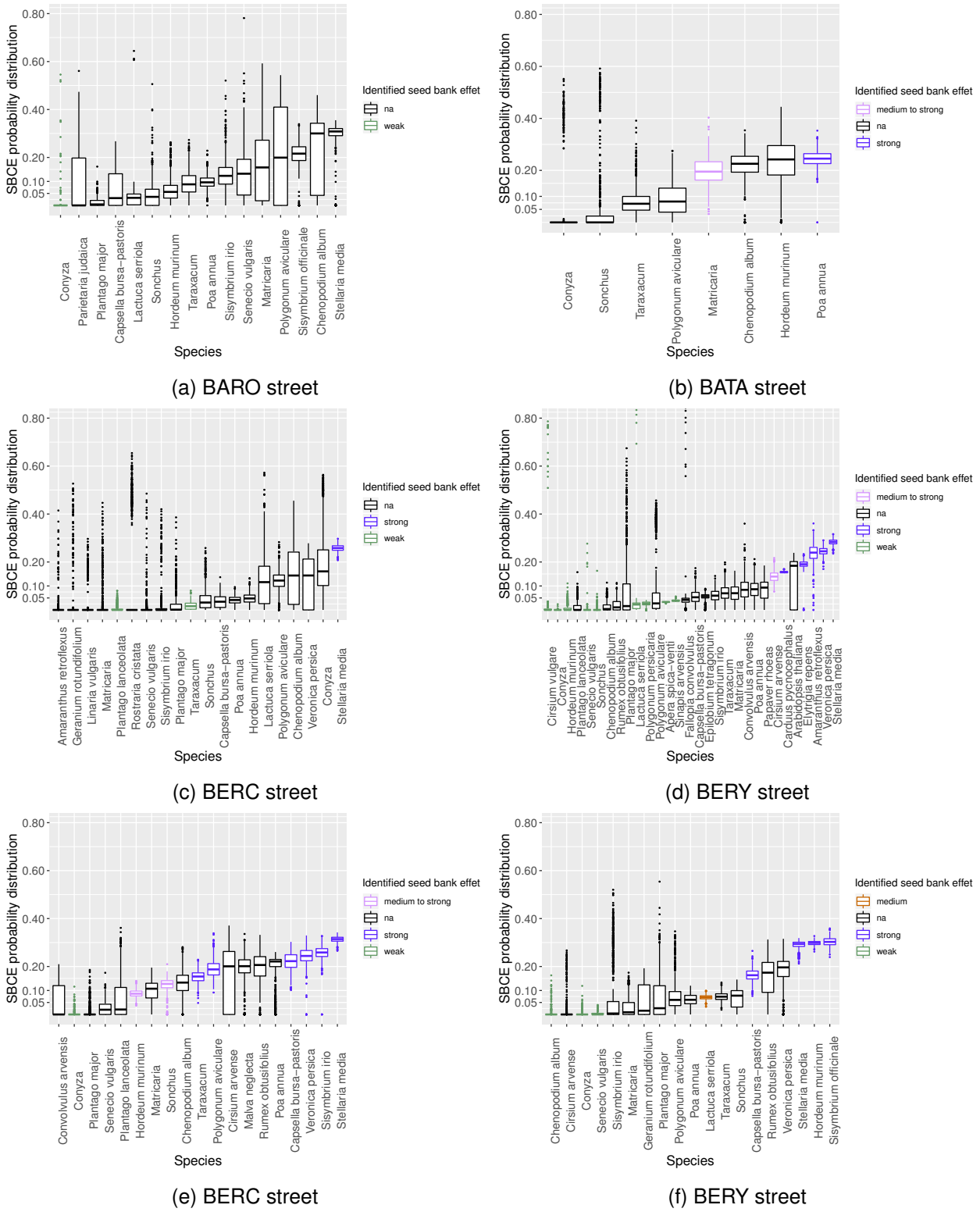
Results of the analysis - Parameter difference between streets

Table 5.11: For each species present in at least 20 tree bases of 2 different streets, AIC of parameter fits assuming metapopulation parameters are respectively different (*AIC- Diff*) or identical (*AIC - Merged*) from one street to another. w_{fusion} stands for the Akaike weight of the model with the same metapopulation parameters for all streets. For each species, we considered as null hypothesis having the same metapopulation parameters for all streets. We used the Holm-Bonferroni method with a 0.05 significance level to account for simultaneous testing. Species in bold are species for which the null hypothesis could not be rejected.

Species name	**AIC - Diff**	**AIC - Merged**	w_{fusion}
Sonchus	14600	17282	$< 10^{-5}$
Conyza	18575	19852	$< 10^{-5}$
Poa annua	17010	18061	$< 10^{-5}$
Veronica persica	3913	4857	$< 10^{-5}$
Matricaria	6027	6758	$< 10^{-5}$
Taraxacum	13787	14385	$< 10^{-5}$
Sisymbrium irio	7980	8534	$< 10^{-5}$
Hordeum murinum	10529	11076	$< 10^{-5}$
Fallopia convolvulus	821	1209	$< 10^{-5}$
Stellaria media	11671	11940	$< 10^{-5}$
Rumex obtusifolius	1402	1658	$< 10^{-5}$
Convolvulus arvensis	959	1214	$< 10^{-5}$
Cirsium arvense	2051	2239	$< 10^{-5}$
Geranium rotundifolium	681	865	$< 10^{-5}$
Capsella bursa-pastoris	7218	7378	$< 10^{-5}$
Polygonum persicaria	602	700	$< 10^{-5}$
Polygonum aviculare	6380	6461	$< 10^{-5}$
Chenopodium album	4373	4439	$< 10^{-5}$
Parietaria judaica	1223	1265	$< 10^{-5}$
Senecio vulgaris	5399	5432	$< 10^{-5}$
Amaranthus retroflexus	853	884	$< 10^{-5}$
Lactuca serriola	2908	2931	1.01×10^{-5}
Plantago major	3361	3380	7.48×10^{-5}
Sisymbrium officinale	1228	1244	3.35×10^{-4}
Plantago lanceolata	1014	1027	1.5×10^{-3}
Cerastium glomeratum	**902**	**902**	**0.5**
Galium aparine	**578**	**575**	**0.622**
Lactuca muralis	**482**	**481**	**0.818**
Sedum vulgare	**633**	**626**	**0.881**
Senecio inaequidens	**2595**	**2591**	**0.971**

Results of the analysis - Distribution of estimated SBCE probabilities

The following figures represent, for each pair or species and street included in the analysis, the distribution of the estimated SBCE probability generated by bootstrap. The boundaries of the boxes indicate the 25th (Q_1) and 75th (Q_3) percentiles, the line in each box indicates the median, and the whiskers indicate $Q_1 - 1.5 \times (Q_3 - Q_1)$ and $Q_3 + 1.5 \times (Q_3 - Q_1)$ for the lower whisker and upper whisker, respectively. The outlying dots show values exceeding this range. The colored boxes indicate the pairs of species and street for which the 95% confidence interval was completely included in $[0, 0.05]$ (*weak seed bank effect*), $[0.05, 1]$ (*medium to strong seed bank effect*), $[0.05, 0.1]$ (*medium seed bank effect*) or $[0.1, 1]$ (*strong seed bank effect*). The signification of the abbreviations used for street names can be found in Table 5.9.



(a) BARO street

(b) BATA street

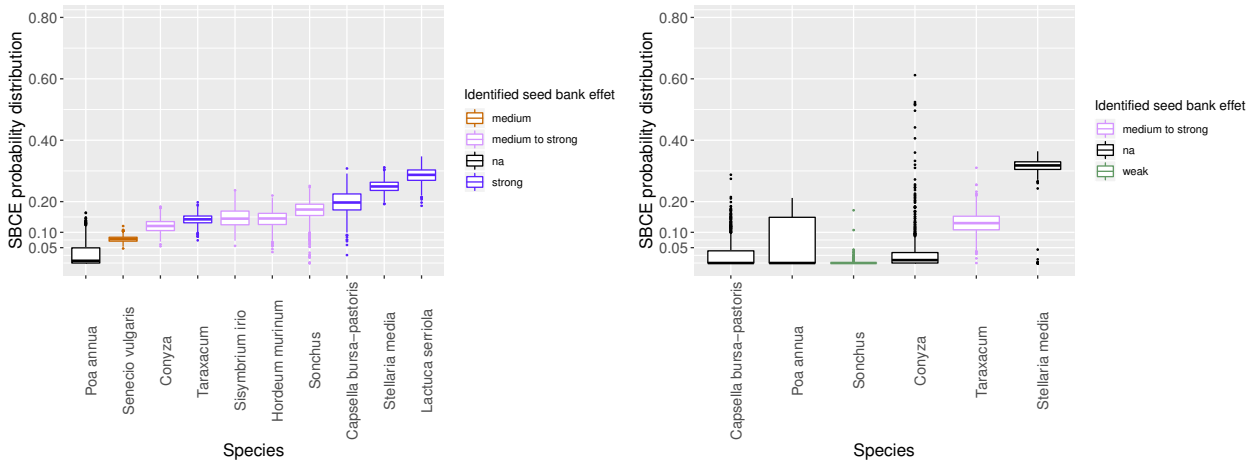
(c) BERC street

(d) BERY street

(e) BERC street

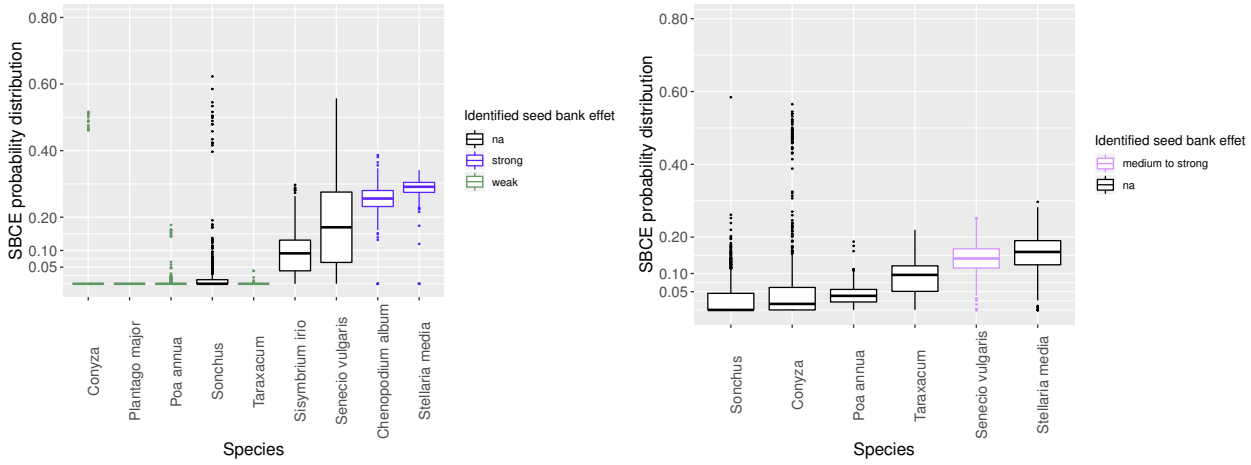
(f) BERY street

Figure 5.10: Distribution of the estimated SBCE probability in the BARO, BATA, BERC, BERY, CHAR and DAUM streets.



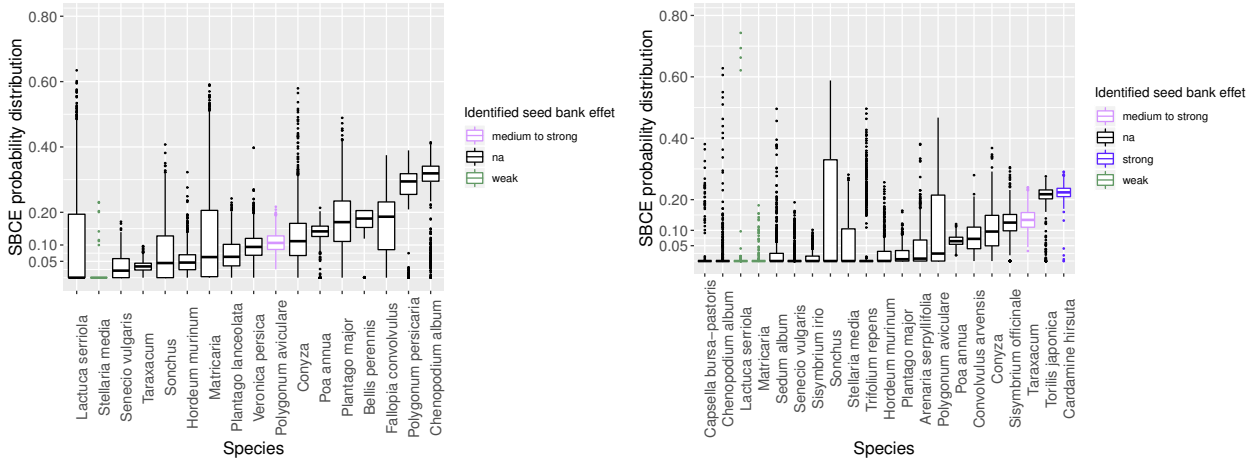
(a) BARO street

(b) BATA street



(c) BERC street

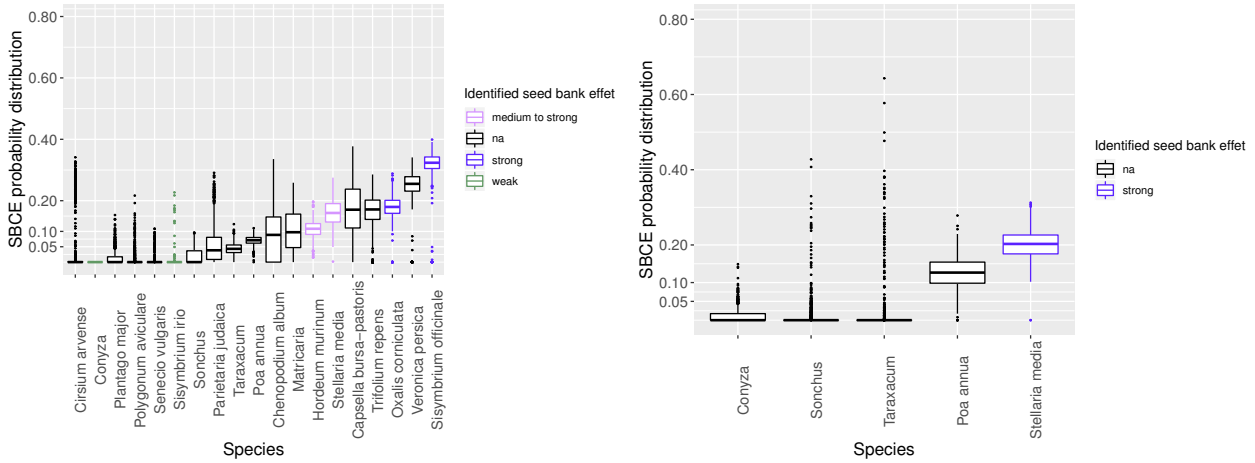
(d) BERY street



(e) BERC street

(f) BERY street

Figure 5.11: Distribution of the estimated SBCE probability in the KESS, LACH, MONT, POMM, RAPE and RBER streets.



(a) BARO street (b) BATA street

Figure 5.12: Distribution of the estimated SBCE probability in the REUI, TAIN streets.

Results of the analysis - Influence of environmental variables and species traits

Because almost all species flower in summer months (June, July and August) and due to high correlation in flowering for consecutive months, we limited the flowering period data to two binary variables : early flowering (months of March, April or May) and late flowering (September or October). We used mixed-effect linear regression model with SBCE probability as response variable, species and street identity as random factors, the type of the closest green space and the dispersal mechanism as fixed qualitative variables, the releasing height of the seeds and the seed weight as fixed quantitative variables and the early and late flowering as binary variables, assuming a gaussian distribution. We checked whether the requirements of independence of residuals, normality of residuals and homogeneity of variances were met. The model was implemented with the lmer function of the lme4 R package [Bat+14]. To take into account the fact that the estimated SBCE probability was not calculated with the same degree of confidence according to species and street, we also integrated into the regression model a weighting vector associated with each value of SBCE probability. These weights W_i were calculated using the formula $W_i = \log(1/(CI{+}_i - CI{-}_i))$, where $CI{+}_i$ and $CI{-}_i$ were respectively the upper and lower bounds of the 95% confidence interval.
Models including and not including these weights both yielded non-significant results. Results presented below (Table 5.12) include the weighting vector.

Table 5.12: Summary of the mixed effects regression of the SBCE probability on species and environmental covariates. For the type of closest green space, estimates are expressed relative to the "B" green space (see Table 5.9 above). For the dispersal mechanism, estimates are expressed relative to the anemochorous dispersal mechanism.

Variable	Estimate	Standard error	t-value	p-value
Green space : type F	-0.038	0.036	-1.046	0.318
Green space : type R	-0.040	0.033	-1.230	0.242
Green space : type S	-0.039	0.039	-1.008	0.336
Releasing height	0.009	0.067	0.139	0.890
Seed weight	-0.005	0.005	-0.872	0.390
Dispersal : autochorous	0.016	0.072	0.217	0.829
Dispersal : barochorous	-0.007	0.034	-0.202	0.842
Dispersal : epizoochorous	0.034	0.041	0.830	0.411
Early flowering	0.040	0.036	1.116	0.271
Late flowering	-0.009	0.034	-0.252	0.803

Results of the analysis - Complete estimation results

Table 5.13: SBCE probability and germination probability estimates for species present in at least 20 tree bases of the street. $CI{-}_i$ and $CI{+}_i$ indicate the lower and upper bounds of the 95% confidence intervals on the estimations of SBCE and germination probabilities, obtained by bootstrap. Abbreviations used for street names are defined in Table 5.9, except for 'NA' which indicates that the estimation parameter were not significantly different from one street to another.

Species	Street	$\mathbb{P}_{SBCE}$	$\mathbb{P}_{SBCE}$ - $CI{-}_i$	$\mathbb{P}_{SBCE}$ - $CI{+}_i$	g	g - $CI{-}_i$	g - $CI{+}_i$
Amaranthus retroflexus	BERC	0	0	0.16	0.473	0.33	0.561
Amaranthus retroflexus	BERY	0.241	0.155	0.298	0.416	0.33	0.517
Apera spica-venti	BERY	0.033	0.03	0.035	0.135	0.13	0.136
Arabidopsis thaliana	BERY	0.202	0	0.22	0.23	0.203	0.445
Arenaria serpyllifolia	RBER	0.053	0	0.172	0.148	0.133	0.205
Bellis perennis	RAPE	0.181	0	0.205	0.156	0.125	0.161
Capsella bursa-pastoris	BARO	0	0	0.196	0.5	0.417	0.575
Capsella bursa-pastoris	BERC	0.033	0	0.095	0.462	0.386	0.532

Table 5.13: SBCE probability and germination probability estimates for species present in at least 20 tree bases of the street. $CI-_i$ and $CI+_i$ indicate the lower and upper bounds of the 95% confidence intervals on the estimations of SBCE and germination probabilities, obtained by bootstrap. Abbreviations used for street names are defined in Table 5.9, except for 'NA' which indicates that the estimation parameter were not significantly different from one street to another.

Species	Street	$\mathbb{P}_{SBCE}$	$\mathbb{P}_{SBCE}$ - $CI-_i$	$\mathbb{P}_{SBCE}$ - $CI+_i$	g	g - $CI-_i$	g - $CI+_i$
Capsella bursa-pastoris	BERY	0.055	0.002	0.115	0.473	0.388	0.544
Capsella bursa-pastoris	CHAR	0.223	0.141	0.282	0.534	0.464	0.602
Capsella bursa-pastoris	DAUM	0.165	0.117	0.213	0.573	0.515	0.634
Capsella bursa-pastoris	KESS	0.199	0.12	0.27	0.625	0.529	0.716
Capsella bursa-pastoris	LACH	0	0	0.149	0.671	0.558	0.756
Capsella bursa-pastoris	RBER	0	0	0.17	0.416	0.318	0.488
Capsella bursa-pastoris	REUI	0.168	0.006	0.353	0.507	0.325	0.624
Cardamine hirsuta	RBER	0.224	0.182	0.266	0.184	0.164	0.211
Carduus pycnocephalus	BERY	0.158	0.155	0.163	0.151	0.151	0.152
Cerastium glomeratum	NA	0.298	0.267	0.32	0.324	0.29	0.375
Chenopodium album	BARO	0.334	0	0.399	0.416	0.335	0.563
Chenopodium album	BATA	0.229	0	0.298	0.364	0.306	0.436
Chenopodium album	BERC	0.209	0	0.37	0.438	0.367	0.571
Chenopodium album	BERY	0.004	0	0.055	0.495	0.452	0.542
Chenopodium album	CHAR	0.131	0.02	0.232	0.349	0.27	0.43
Chenopodium album	DAUM	0	0	0.036	0.245	0.156	0.316
Chenopodium album	MONT	0.259	0.182	0.33	0.289	0.217	0.375
Chenopodium album	RAPE	0.325	0.037	0.385	0.32	0.247	0.421
Chenopodium album	RBER	0	0	0.135	0.207	0.118	0.333
Chenopodium album	REUI	0	0	0.227	0.385	0.28	0.45
Cirsium arvense	BERY	0.143	0.096	0.186	0.338	0.259	0.443
Cirsium arvense	CHAR	0.219	0	0.332	0.223	0.158	0.316
Cirsium arvense	DAUM	0	0	0.202	0.301	0.129	0.486
Cirsium arvense	REUI	0.182	0	0.291	0.222	0.166	0.47
Cirsium vulgare	BERY	0	0	0.034	0.122	0.116	0.157
Convolvulus arvensis	BERY	0.084	0	0.171	0.561	0.425	0.665
Convolvulus arvensis	CHAR	0	0	0.183	0.14	0.13	0.177
Convolvulus arvensis	RBER	0.074	0	0.177	0.639	0.558	0.704
Conyza	BARO	0	0	0	0.382	0.292	0.457
Conyza	BATA	0	0	0.451	0.489	0.296	0.618
Conyza	BERC	0.164	0.018	0.524	0.39	0.338	0.433
Conyza	BERY	0	0	0	0.487	0.448	0.523
Conyza	CHAR	0	0	0.023	0.264	0.203	0.333
Conyza	DAUM	0	0	0.021	0.269	0.228	0.323
Conyza	KESS	0.121	0.084	0.162	0.214	0.184	0.249
Conyza	LACH	0.013	0	0.203	0.484	0.373	0.586
Conyza	MONT	0	0	0	0.529	0.484	0.562
Conyza	POMM	0.024	0	0.478	0.458	0.388	0.518
Conyza	RAPE	0.114	0.004	0.369	0.405	0.352	0.451
Conyza	RBER	0.096	0	0.257	0.438	0.386	0.487
Conyza	REUI	0	0	0	0.414	0.362	0.467
Conyza	TAIN	0	0	0.063	0.382	0.307	0.451
Elytrigia repens	BERY	0.192	0.168	0.213	0.185	0.164	0.209
Epilobium tetragonum	BERY	0.058	0	0.076	0.145	0.13	0.168
Fallopia convolvulus	BERY	0.043	0	0.06	0.175	0.154	0.197
Fallopia convolvulus	RAPE	0.196	0	0.292	0.336	0.258	0.446
Galium aparine	NA	0.304	0.014	0.315	0.297	0.253	0.485
Geranium rotundifolium	BERC	0	0	0.366	0.397	0.251	0.519
Geranium rotundifolium	DAUM	0.094	0	0.168	0.162	0.133	0.267
Hordeum murinum	BARO	0.056	0	0.181	0.692	0.623	0.744
Hordeum murinum	BATA	0.251	0	0.396	0.43	0.346	0.52
Hordeum murinum	BERC	0.047	0.008	0.091	0.689	0.641	0.728
Hordeum murinum	BERY	0	0	0.047	0.617	0.568	0.659
Hordeum murinum	CHAR	0.088	0.056	0.119	0.733	0.689	0.771
Hordeum murinum	DAUM	0.299	0.281	0.314	0.431	0.401	0.458
Hordeum murinum	KESS	0.148	0.081	0.19	0.776	0.723	0.834
Hordeum murinum	RAPE	0.047	0	0.137	0.528	0.474	0.576
Hordeum murinum	RBER	0	0	0.133	0.315	0.199	0.406
Hordeum murinum	REUI	0.108	0.05	0.154	0.733	0.686	0.787
Lactuca muralis	NA	0.207	0.213	0.298	0.174	0.166	0.194
Lactuca serriola	BARO	0.032	0	0.077	0.246	0.169	0.354

Table 5.13: SBCE probability and germination probability estimates for species present in at least 20 tree bases of the street. $CI-_i$ and $CI+_i$ indicate the lower and upper bounds of the 95% confidence intervals on the estimations of SBCE and germination probabilities, obtained by bootstrap. Abbreviations used for street names are defined in Table 5.9, except for 'NA' which indicates that the estimation parameter were not significantly different from one street to another.

Species	Street	$\mathbb{P}_{SBCE}$	$\mathbb{P}_{SBCE}$ - $CI-_i$	$\mathbb{P}_{SBCE}$ - $CI+_i$	g	g - $CI-_i$	g - $CI+_i$
Lactuca serriola	BERC	0.138	0	0.344	0.22	0.148	0.297
Lactuca serriola	BERY	0.026	0	0.039	0.148	0.136	0.162
Lactuca serriola	DAUM	0.072	0.052	0.091	0.136	0.126	0.149
Lactuca serriola	KESS	0.289	0.23	0.329	0.291	0.229	0.366
Lactuca serriola	RAPE	0	0	0.498	0.152	0.136	0.264
Lactuca serriola	RBER	0	0	0.011	0.139	0.125	0.182
Linaria vulgaris	BERC	0	0	0.193	0.206	0.136	0.32
Malva neglecta	CHAR	0.208	0.003	0.281	0.195	0.165	0.227
Matricaria	BARO	0.189	0	0.463	0.342	0.268	0.43
Matricaria	BATA	0.194	0.099	0.306	0.509	0.434	0.574
Matricaria	BERC	0	0	0.26	0.393	0.241	0.499
Matricaria	BERY	0.07	0	0.138	0.695	0.64	0.745
Matricaria	CHAR	0.116	0	0.17	0.179	0.155	0.223
Matricaria	DAUM	0.006	0	0.129	0.336	0.211	0.422
Matricaria	RAPE	0.067	0	0.513	0.294	0.202	0.374
Matricaria	RBER	0	0	0.019	0.384	0.237	0.485
Matricaria	REUI	0.095	0	0.223	0.313	0.199	0.452
Oxalis corniculata	REUI	0.18	0.108	0.244	0.214	0.167	0.268
Papaver rhoeas	BERY	0.115	0	0.152	0.252	0.18	0.527
Parietaria judaica	BARO	0	0	0.312	0.466	0.344	0.572
Parietaria judaica	REUI	0.039	0	0.232	0.502	0.289	0.611
Plantago lanceolata	BERC	0	0	0.04	0.509	0.333	0.629
Plantago lanceolata	BERY	0	0	0.057	0.368	0.19	0.5
Plantago lanceolata	CHAR	0.008	0	0.276	0.163	0.112	0.333
Plantago lanceolata	RAPE	0.066	0	0.174	0.517	0.411	0.603
Plantago major	BARO	0.006	0	0.055	0.564	0.417	0.665
Plantago major	BERC	0.004	0	0.141	0.368	0.256	0.463
Plantago major	BERY	0.077	0	0.461	0.225	0.155	0.339
Plantago major	CHAR	0	0	0.09	0.165	0.115	0.277
Plantago major	DAUM	0.032	0	0.279	0.323	0.244	0.396
Plantago major	MONT	0	0	0	0.351	0.231	0.459
Plantago major	RAPE	0.169	0.033	0.393	0.294	0.223	0.372
Plantago major	RBER	0.013	0	0.081	0.282	0.199	0.361
Plantago major	REUI	0	0	0.08	0.517	0.419	0.585
Poa annua	BARO	0.097	0.047	0.15	0.795	0.751	0.831
Poa annua	BATA	0.245	0.191	0.305	0.604	0.542	0.658
Poa annua	BERC	0.043	0.003	0.077	0.731	0.697	0.771
Poa annua	BERY	0.086	0.012	0.164	0.647	0.59	0.697
Poa annua	CHAR	0.227	0	0.243	0.627	0.603	0.675
Poa annua	DAUM	0.062	0.018	0.107	0.808	0.783	0.83
Poa annua	KESS	0.012	0	0.127	0.734	0.687	0.772
Poa annua	LACH	0.171	0	0.19	0.829	0.797	0.866
Poa annua	MONT	0	0	0.019	0.781	0.741	0.814
Poa annua	POMM	0.044	0	0.093	0.834	0.798	0.872
Poa annua	RAPE	0.146	0	0.188	0.728	0.683	0.772
Poa annua	RBER	0.064	0.032	0.1	0.751	0.723	0.782
Poa annua	REUI	0.071	0.039	0.094	0.854	0.83	0.875
Poa annua	TAIN	0.128	0.04	0.209	0.606	0.55	0.67
Polygonum aviculare	BARO	0.418	0	0.495	0.327	0.281	0.42
Polygonum aviculare	BATA	0.085	0	0.224	0.584	0.528	0.652
Polygonum aviculare	BERC	0.126	0.01	0.206	0.53	0.472	0.59
Polygonum aviculare	BERY	0.03	0	0.406	0.399	0.289	0.477
Polygonum aviculare	CHAR	0.188	0.126	0.273	0.436	0.387	0.481
Polygonum aviculare	DAUM	0.064	0.006	0.256	0.479	0.391	0.563
Polygonum aviculare	RAPE	0.106	0.055	0.176	0.637	0.578	0.688
Polygonum aviculare	RBER	0.021	0	0.389	0.342	0.218	0.469
Polygonum aviculare	REUI	0	0	0.092	0.514	0.442	0.575
Polygonum persicaria	BERY	0.027	0	0.037	0.129	0.128	0.139
Polygonum persicaria	RAPE	0.312	0	0.36	0.339	0.288	0.412
Rostraria cristata	BERC	0	0	0.551	0.31	0.244	0.393
Rumex obtusifolius	BERY	0.006	0	0.069	0.471	0.18	0.599

Table 5.13: SBCE probability and germination probability estimates for species present in at least 20 tree bases of the street. $CI-_i$ and $CI+_i$ indicate the lower and upper bounds of the 95% confidence intervals on the estimations of SBCE and germination probabilities, obtained by bootstrap. Abbreviations used for street names are defined in Table 5.9, except for 'NA' which indicates that the estimation parameter were not significantly different from one street to another.

Species	Street	$\mathbb{P}_{SBCE}$	$\mathbb{P}_{SBCE}$ - $CI-_i$	$\mathbb{P}_{SBCE}$ - $CI+_i$	g	g - $CI-_i$	g - $CI+_i$
Rumex obtusifolius	CHAR	0.216	0	0.3	0.213	0.167	0.343
Rumex obtusifolius	DAUM	0.177	0	0.272	0.271	0.203	0.436
Sedum album	RBER	0	0	0.178	0.289	0.204	0.38
Sedum vulgare	NA	0.214	0.203	0.23	0.174	0.165	0.192
Senecio inaequidens	NA	0.001	0	0.051	0.16	0.143	0.195
Senecio vulgaris	BARO	0.15	0	0.309	0.208	0.143	0.29
Senecio vulgaris	BERC	0	0	0.109	0.17	0.12	0.247
Senecio vulgaris	BERY	0	0	0	0.259	0.196	0.315
Senecio vulgaris	CHAR	0.02	0	0.089	0.24	0.159	0.304
Senecio vulgaris	DAUM	0	0	0.025	0.165	0.145	0.2
Senecio vulgaris	KESS	0.078	0.058	0.098	0.181	0.155	0.212
Senecio vulgaris	MONT	0.201	0	0.441	0.238	0.166	0.35
Senecio vulgaris	POMM	0.143	0.057	0.213	0.185	0.143	0.233
Senecio vulgaris	RAPE	0.027	0	0.114	0.228	0.2	0.287
Senecio vulgaris	RBER	0	0	0.109	0.174	0.154	0.223
Senecio vulgaris	REUI	0	0	0.063	0.262	0.166	0.348
Sinapis arvensis	BERY	0.039	0.035	0.046	0.138	0.131	0.153
Sisymbrium irio	BARO	0.119	0	0.265	0.444	0.302	0.56
Sisymbrium irio	BERC	0	0	0.257	0.617	0.519	0.671
Sisymbrium irio	BERY	0.061	0.015	0.109	0.636	0.577	0.686
Sisymbrium irio	CHAR	0.261	0.201	0.303	0.275	0.224	0.331
Sisymbrium irio	DAUM	0	0	0.397	0.327	0.254	0.402
Sisymbrium irio	KESS	0.146	0.092	0.206	0.333	0.271	0.401
Sisymbrium irio	MONT	0.091	0	0.215	0.461	0.324	0.553
Sisymbrium irio	RBER	0	0	0.06	0.348	0.182	0.482
Sisymbrium irio	REUI	0	0	0	0.559	0.458	0.633
Sisymbrium officinale	BARO	0.225	0	0.298	0.215	0.166	0.296
Sisymbrium officinale	DAUM	0.304	0.264	0.338	0.219	0.188	0.254
Sisymbrium officinale	RBER	0.128	0.019	0.21	0.443	0.308	0.57
Sisymbrium officinale	REUI	0.326	0.192	0.375	0.24	0.197	0.35
Sonchus	BARO	0.042	0	0.161	0.393	0.315	0.475
Sonchus	BATA	0.003	0	0.499	0.405	0.214	0.553
Sonchus	BERC	0.035	0	0.17	0.416	0.338	0.48
Sonchus	BERY	0	0	0.02	0.386	0.329	0.438
Sonchus	CHAR	0.127	0.056	0.166	0.188	0.166	0.212
Sonchus	DAUM	0.078	0	0.127	0.178	0.158	0.237
Sonchus	KESS	0.177	0.078	0.22	0.284	0.246	0.363
Sonchus	LACH	0	0	0.002	0.372	0.265	0.463
Sonchus	MONT	0	0	0.133	0.35	0.29	0.398
Sonchus	POMM	0	0	0.157	0.202	0.177	0.246
Sonchus	RAPE	0.048	0	0.261	0.523	0.451	0.578
Sonchus	RBER	0.371	0	0.531	0.324	0.28	0.412
Sonchus	REUI	0.011	0	0.076	0.257	0.196	0.346
Sonchus	TAIN	0	0	0.094	0.269	0.184	0.351
Stellaria media	BARO	0.316	0	0.341	0.33	0.276	0.417
Stellaria media	BERC	0.259	0.229	0.284	0.397	0.355	0.442
Stellaria media	BERY	0.284	0.258	0.305	0.318	0.276	0.363
Stellaria media	CHAR	0.315	0.286	0.335	0.325	0.289	0.365
Stellaria media	DAUM	0.298	0.243	0.311	0.487	0.448	0.542
Stellaria media	KESS	0.252	0.213	0.287	0.347	0.302	0.395
Stellaria media	LACH	0.325	0	0.35	0.421	0.309	0.544
Stellaria media	MONT	0.292	0.242	0.327	0.343	0.279	0.414
Stellaria media	POMM	0.162	0.04	0.246	0.379	0.25	0.521
Stellaria media	RAPE	0	0	0	0.476	0.424	0.53
Stellaria media	RBER	0	0	0.232	0.451	0.379	0.512
Stellaria media	REUI	0.16	0.081	0.244	0.625	0.559	0.679
Stellaria media	TAIN	0.208	0.128	0.275	0.273	0.221	0.339
Taraxacum	BARO	0.088	0.01	0.205	0.621	0.567	0.667
Taraxacum	BATA	0.075	0	0.203	0.425	0.316	0.521
Taraxacum	BERC	0.016	0	0.048	0.626	0.572	0.678
Taraxacum	BERY	0.075	0	0.135	0.547	0.506	0.59

Table 5.13: SBCE probability and germination probability estimates for species present in at least 20 tree bases of the street. $CI-_i$ and $CI+_i$ indicate the lower and upper bounds of the 95% confidence intervals on the estimations of SBCE and germination probabilities, obtained by bootstrap. Abbreviations used for street names are defined in Table 5.9, except for 'NA' which indicates that the estimation parameter were not significantly different from one street to another.

Species	Street	$\mathbb{P}_{SBCE}$	$\mathbb{P}_{SBCE}$ - $CI-_i$	$\mathbb{P}_{SBCE}$ - $CI+_i$	g	g - $CI-_i$	g - $CI+_i$
Taraxacum	CHAR	0.157	0.107	0.204	0.64	0.592	0.688
Taraxacum	DAUM	0.075	0.042	0.111	0.715	0.682	0.748
Taraxacum	KESS	0.147	0.103	0.174	0.748	0.695	0.796
Taraxacum	LACH	0.134	0.062	0.197	0.617	0.53	0.697
Taraxacum	MONT	0	0	0	0.314	0.239	0.377
Taraxacum	POMM	0.116	0	0.174	0.203	0.148	0.356
Taraxacum	RAPE	0.035	0.004	0.066	0.814	0.775	0.851
Taraxacum	RBER	0.135	0.068	0.209	0.581	0.527	0.629
Taraxacum	REUI	0.044	0.01	0.082	0.668	0.62	0.709
Taraxacum	TAIN	0	0	0.233	0.188	0.143	0.295
Torilis japonica	RBER	0.089	0.005	0.257	0.526	0.201	0.42
Trifolium repens	RBER	0.221	0	0.327	0.249	0.152	0.552
Trifolium repens	REUI	0.176	0	0.261	0.182	0.147	0.282
Veronica persica	BERC	0.148	0	0.258	0.336	0.243	0.501
Veronica persica	BERY	0.246	0.212	0.275	0.309	0.261	0.358
Veronica persica	CHAR	0.246	0.169	0.301	0.207	0.153	0.268
Veronica persica	DAUM	0.208	0	0.264	0.243	0.196	0.456
Veronica persica	RAPE	0.017	0.017	0.191	0.208	0.476	0.645
Veronica persica	RAPE	0.094	0.017	0.191	0.569	0.476	0.645
Veronica persica	REUI	0.262	0	0.317	0.238	0.185	0.426

5.5.5 Tutorial on SBCE probability estimation

The codes to perform SBCE probability estimation along with an example data set (*test.csv*) are available on a GitHub repository :

https://github.com/apollinelouvet/sbce_probability.git

The codes work with Python 3.3 or more recent versions.

Data formatting

The program takes as input a presence/absence dataset in the csv format with each line corresponding to a year of observation (in chronological order) and each column corresponding to a patch. Refer to the provided 'test.csv' file for an example. The presence of the focus species is indicated by the entry 1 and the absence by 0. If a year of observation is missing, the corresponding line in the csv file should be filled with zeroes. The first two lines of the csv file are not read by the codes and should be filled with zeros. Therefore, the first year of observation corresponds to the third line of the csv file.

SBCE probability estimation

The codes

quick_code_without_missing_years.py

and

quick_code_with_one_missing_year.py

can be used to estimate the SBCE probability, respectively when no observation is missing and when one year of observation is missing. The name of the file to be analysed has to be indicated in the code, as a character string not containing the .csv extension.

When one year of observation is missing, this year can be indicated in the code as well, the first year of observation being numbered as 0.
The algorithm returns the estimated SBCE probability, along with the AIC on PRM parameters estimation. The execution time of the algorithm for the number of patches and monitoring durations we considered ranges from a few seconds to several minutes.

95 % confidence interval on SBCE probability estimation
The codes

code_without_missing_years.py

and

code_with_one_missing_year.py

can be used to compute a 95% confidence interval on SBCE probability estimation, respectively when no observation is missing and when one year of observation is missing. The algorithm returns the estimated SBCE probability, the AIC on PRM parameters estimation, and a csv file containing SBCE probability estimates for 1000 datasets generated by bootstrap. This csv file can be used to analyse the distribution of the estimated SBCE probability, and in particular to get the 95% confidence interval on SBCE probability estimation. The execution time of the algorithm can take up to several hours.

Chapter 6

Extinction threshold and large population limit of a plant metapopulation model with recurrent extinction events and a seed bank component

This chapter is based on the article [Lou22], available on Arxiv and accepted for publication in Theoretical Population Biology.

Abstract

We introduce a new model for plant metapopulations with a seed bank component, living in a fragmented environment in which local extinction events are frequent. This model is an intermediate between population dynamics models with a seed bank component, based on the classical Wright-Fisher model, and Stochastic Patch Occupancy Models (SPOMs) used in metapopulation ecology. Its main feature is the use of "ghost" individuals, which can reproduce but with a very strong selective disadvantage against "real" individuals, to artificially ensure a constant population size. We show the existence of an extinction threshold above which persistence of the subpopulation of "real" individuals is not possible, and investigate how the seed bank characteristics affect this extinction threshold. We also show the convergence of the model to a SPOM under an appropriate scaling, bridging the gap between individual-based models and occupancy models.

6.1 Introduction

Understanding how plant populations survive in fragmented landscapes is an important question in ecology and conservation biology [Fah03]. One potential driver of plant populations' persistence is the ability to form a seed bank, which greatly influences population and community dynamics [Fen95]. For such plant species, the seeds produced can stay dormant in the soil for up to several decades depending on the species, without losing viability [BB14]. See [Len+21] for an overview of seed bank characteristics and properties, along with the emergent phenomena it can generate.

Populations living in fragmented landscapes are often modelled as metapopulations, that is, as populations distributed over a set of interconnected patches. Metapopulations are also frequently characterized by recurrent local extinction events, regional persistence being the result of a balance between colonization (from neighbouring patches or from an external source) and local extinction events [Lev69; MW67]. See [HGM97] for a general introduction to metapopulation theory.

Many classical metapopulation models, such as the Levins model [Lev69] or the Propagule Rain model [Got91], describe the occupancy of each patch (i.e whether the species of interest is present or absent in each of the patches) and do not depend on, nor model, the actual census numbers. These models are referred to as Stochastic Patch Occupancy models, or SPOMs. Since presence/absence data is easier to collect than abundance data, and since parameter inference is possible for a broad range of SPOMs (see e.g [Moi99; Moi04; Plu+18]), they are well-suited to the study of real metapopulations. Classical metapopulation models do not account for seed dormancy, but more recently models incorporating a seed bank component were also developed [Bor+15; Fré+13; Plu+18]. The model introduced in [Plu+18] was successfully applied to plant metapopulations in highly disturbed environments, such as weeds in agroecosystems [Plu+18] or plants in urban tree bases [Lou+21], highlighting that some plant species monitored did have a seed bank.

In population genetics, metapopulation models often describe the number and genetic types of individuals rather than the occupancy in each patch. They are usually defined by first specifying an intra-patch dynamic, and then adding migration between patches. The migration process can heavily depend on the underlying geographical structure, as in the stepping-stone model [KW64], or not depend on it at all, as in Wright's island model [Wri31]. See e.g [LM15; TV09; WA01] and references therein for examples of metapopulation models based on Wright's island model, and [Aus+97; Bar+13; PE15] and references therein for examples of metapopulation models based on the stepping-stone model.

Models used to specify the intra-patch dynamic can be classical population dynamics models, without any intra-patch spatial structure, provided patches are considered as sufficiently small to neglect spatial effects in each one of them. The geographical structure in the metapopulation model is then only contained in the localization of the patches. The intra-patch dynamic can comprise a seed bank, using population dynamics models with a seed bank component, such as the ones based on the Wright-Fisher model. In the original Wright-Fisher model, the population size (in a single patch) is constant through time and equal to N, and each individual has a genetic type, or allele. In each generation, each one of the N new individuals chooses a parent uniformly at random among the N individuals in the previous generation, and adopts its type. Including a seed bank in the Wright-Fisher model implies choosing a parent potentially not in the previous generation, but at least two generations ago, the maximal number of potentially contributing generations being bounded [KKL01] or not [Bla+13; Bla+16]. See [BK20] for a review of seed bank models in population genetics, and [HP17; GHO22; ŽT12] for extensions of the Wright-Fisher model with a seed bank component to metapopulations.

For plant metapopulations in which extinction events are frequent, we can expect the population size of each patch to vary a lot from one generation to the next. This contradicts the constant patch

population size hypothesis underlying the use of a Wright-Fisher model. In order to incorporate extinction event-induced fluctuations in a Wright-Fisher model, it is possible to adopt the approach used in [DF16; HN08], based on a long-range biased voter model: assign a maximal population size to each patch, and fill the remaining space with "ghost", or type 0, individuals. In this framework, each patch contains both type 1 "real" individuals and type 0 "ghost" individuals, the former having a very strong selective advantage over the latter (in the spirit of [Lou21]). As in [DF16; HN08], the model can then be interpreted as an interacting particle system, in which a 0 corresponds to a ghost individual and a 1 to a real individual.

In this chapter, we introduce a new individual-based metapopulation model for plant metapopulations in which local extinction events are frequent. This model is primarily suited to annual plants living in highly disturbed patchy environments, such as urban tree bases or agroecosystems. It is also adapted to other plant species living in such environments, provided each patch is "emptied" at the end of each generation (for instance by gardeners in an urban environment or by farmers in an agroecosystem). The intra-patch dynamics will be based on a variant of a Wright-Fisher model with a seed bank component, using ghost individuals to allow for fluctuating patch population sizes. It will use the model introduced in [Bla+16], with an extra bound introduced on the number of generations a seed can stay dormant without losing viability. Indeed, for some plant species, seeds lose viability after only one or two years of dormancy [BB14]. This is reminiscent of the model introduced in [KKL01], in the sense that real individuals come from parents living a bounded number of generations ago. The main difference is that in our model, individuals first choose parents living potentially arbitrarily far ago in the past, and then obtain their (real or ghost) type depending on their choice.

In order to bridge the gap between individual-based metapopulation models and SPOMs, we will show that our metapopulation process can be embedded in a SPOM. Moreover, we will prove that under an appropriate scaling of patch population size and of the "selection" strength of real individuals against ghost individuals (quantified by the parameter k), the individual-based metapopulation process converges to this SPOM. The convergence result has two applications. First, from a theoretical viewpoint, it shows that a specific SPOM (or presence/absence-based model) is the scaling limit of an individual-based metapopulation model. Then, we will use the convergence result and the embedding in order to show the existence of an extinction threshold for metapopulation persistence, depending only on the seed bank parameters, and highlighting how the presence of a seed bank can prevent metapopulation extinction.

While the metapopulation model we will introduce and study is based on models coming from population genetics, this chapter will not focus on the study of the genetic diversity in such populations, which is deferred to future work. Instead, the aims of this work are threefold:

1. Introduce a general individual-based metapopulation model with a seed bank component, in which local extinction events can be frequent and patch population sizes can vary from one generation to the next.
2. Show the existence of an extinction threshold depending on the seed bank parameters.
3. Bridge the gap between SPOMs and individual-based metapopulation models by showing that in a well-chosen parameter regime, the individual-based metapopulation model we consider converges to a SPOM.

6.1.1 The k-parent Wright-Fisher metapopulation process with seed bank

In the whole paper, we use the notation $\mathbb{N} = \{0, 1, 2, ...\}$, and for all $M \in \mathbb{N}\backslash\{0\}$, $[\![1, M]\!] = \{1, ..., M\}$.

We will consider that the metapopulation is formed by an infinite number of patches arranged in a line. A patch contains a fixed number of *seed bank compartments*, each one containing exactly one seed: either a ghost (type 0) seed, or a real (type 1) seed. In order to define the metapopulation

model, we describe how in each generation, seeds germinate and grow into plants which produce new seeds and die. Concretely, the metapopulation model will only record the composition of the *seed bank* at the beginning of each generation, and not the standing vegetation in each patch in each generation.

In all that follows, let $M \in \mathbb{N}^*$, $H \in \mathbb{N}$, $k \in \mathbb{N}\backslash\{0,1\}$, $g \in (0,1)$, $c \in (0,1/2)$ and $p \in [0,1]$. We assume that $\lfloor gM \rfloor \geq 1$. Patches will be indexed by $i \in \mathbb{Z}$, and seed bank compartments inside a patch by $j \in [\![1, M]\!]$. The notation (i,j) will correspond to the seed bank compartment j in patch i.

The following two spaces will be used to describe the initial types and the age of the seeds occupying the seed bank compartments:

$$\mathcal{F}_M := \left\{ (\xi_{i,j})_{i\in\mathbb{Z}, j\in[\![1,M]\!]} \in \{0,1\}^{\mathbb{Z}\times[\![1,M]\!]} : \mathrm{Card}\left(\{(i,j) \in \mathbb{Z}\times[\![1,M]\!] : \xi_{i,j} = 1\}\right) < +\infty \right\},$$
$$\text{and } \mathcal{H}_M := \left\{ (h_{i,j})_{i\in\mathbb{Z}, j\in[\![1,M]\!]} : \forall (i,j) \in \mathbb{Z}\times[\![1,M]\!], h_{i,j} \in \mathbb{N} \right\}.$$

$(\xi, h) \in \mathcal{F}_M \times \mathcal{H}_M$ corresponds to a metapopulation in which for all $(i,j) \in \mathbb{Z}\times[\![1,M]\!]$, the seed occupying the seed bank compartment (i,j) is of age $h_{i,j}$ and of type $\xi_{i,j}\mathbb{1}_{\{h_{i,j}\leq H\}}$. That is, the seed in (i,j) was originally of type $\xi_{i,j}$ when it was produced, but may have expired since then.

The k-parent Wright-Fisher metapopulation process with seed bank is defined in the following way.

Definition 6.1.1. *(k-parent WFSB metapopulation process) Let $(\xi, h) \in \mathcal{F}_M \times \mathcal{H}_M$. The k-parent Wright-Fisher metapopulation process with seed bank, with parameters (M, H, g, c, p) and initial condition (ξ, h) and denoted by $(\xi^n, h^n)_{n\in\mathbb{N}}$, is the $(\mathcal{F}_M \times \mathcal{H}_M)$-valued Markov chain defined by $(\xi^0, h^0) = (\xi, h)$ and for all $n \in \mathbb{N}$, given (ξ^n, h^n) :*

1. *For each $i \in \mathbb{Z}$, we sample $\lfloor gM \rfloor$ different seed bank compartments $s_{i,1}, ..., s_{i,\lfloor gM\rfloor} \in [\![1, M]\!]$ uniformly at random in patch i.*

2. *Let $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ be i.i.d $\{0,1\}$-valued random variables such that $\mathbb{P}(\mathrm{Ext}_1 = 1) = p$.*

3. *For all $(i,j) \in \mathbb{Z}\times[\![1,M]\!]$, if $j \in \{s_{i,j'} : j' \in [\![1, \lfloor gM\rfloor]\!]\}$, we first set $h^{n+1}_{i,j} = 0$. Moreover, let $C_{1,i,j}, ..., C_{k,i,j}$ be i.i.d $\{-1,0,1\}$-valued random variables such that*

$$\mathbb{P}(C_{1,i,j} = 1) = \mathbb{P}(C_{1,i,j} = -1) = c,$$

and such that $(C_{l,i,j})_{1\leq l\leq k}$ is independent from $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ and from $(C_{i',j',l})_{1\leq l\leq k}$ for $(i',j') \neq (i,j)$.

For all $l \in [\![1,k]\!]$, if $\mathrm{Ext}_{i+C_{l,i,j}} = 1$, we set $\tilde{k}_l = 0$, and if $\mathrm{Ext}_{i+C_{l,i,j}} = 0$, we sample one seed bank compartment j_l uniformly at random among the $\lfloor gM \rfloor$ ones sampled in the patch $i+C_{l,i,j}$ (those in the set $\{s_{i+C_{l,i,j},j'} : j' \in [\![1, \lfloor gM\rfloor]\!]\}$), and we set

$$\tilde{k}_l = \xi^n_{i+C_{l,i,j},j_l}\mathbb{1}_{\{h^n_{i+C_{l,i,j},j_l}\leq H\}}.$$

We conclude by setting $\xi^{n+1}_{i,j} = \max\{\tilde{k}_l : l \in [\![1,k]\!]\}$.

4. *On the other hand, if $j \notin \{s_{i,j'} : j' \in [\![1, \lfloor gM\rfloor]\!]\}$, we set $\xi^{n+1}_{i,j} = \xi^n_{i,j}$ and $h^{n+1}_{i,j} = h^n_{i,j} + 1$.*

Intuitively, the k-parent WFSB metapopulation process evolves as follows.

1. At each generation, exactly $\lfloor gM \rfloor$ seeds germinate in each patch. Type 0 seeds yield (ghost) type 0 plants, while type 1 seeds yield (real) type 1 plants *only if the seed was produced less than $H+1$ generations ago*, i.e, only if it has not expired.

2. Then, each patch is affected by an extinction event independently from other patches and with probability p. During an extinction event, all the juvenile plants in the patch become type 0 plants.

3. In each patch, the $\lfloor gM \rfloor$ empty seed bank compartments are filled with new seeds in the following way. For each compartment, k potential parents are chosen uniformly at random, each one of them being chosen in the same patch with probability $1 - 2c$, or in the patch on the left (resp. on the right) with probability c. The same potential parent may be chosen more than once for the same seed bank compartment. If all the k plants chosen as potential parents are of type 0, then the seed bank compartment is filled with a type 0 seed produced by the last plant chosen. Conversely, if at least one of the k plants chosen is of type 1, then the first type 1 plant chosen produces a seed which fills the seed bank compartment.

4. The remaining seeds stay dormant in the seed bank until the next generation.

See Figure 6.1 for an illustration of this dynamics. As mentioned above, observe that while the dynamics involves seeds germinating, growing into plants which produce new seeds and then die, the model only encodes the *seed bank composition*, and not the types of the plants.

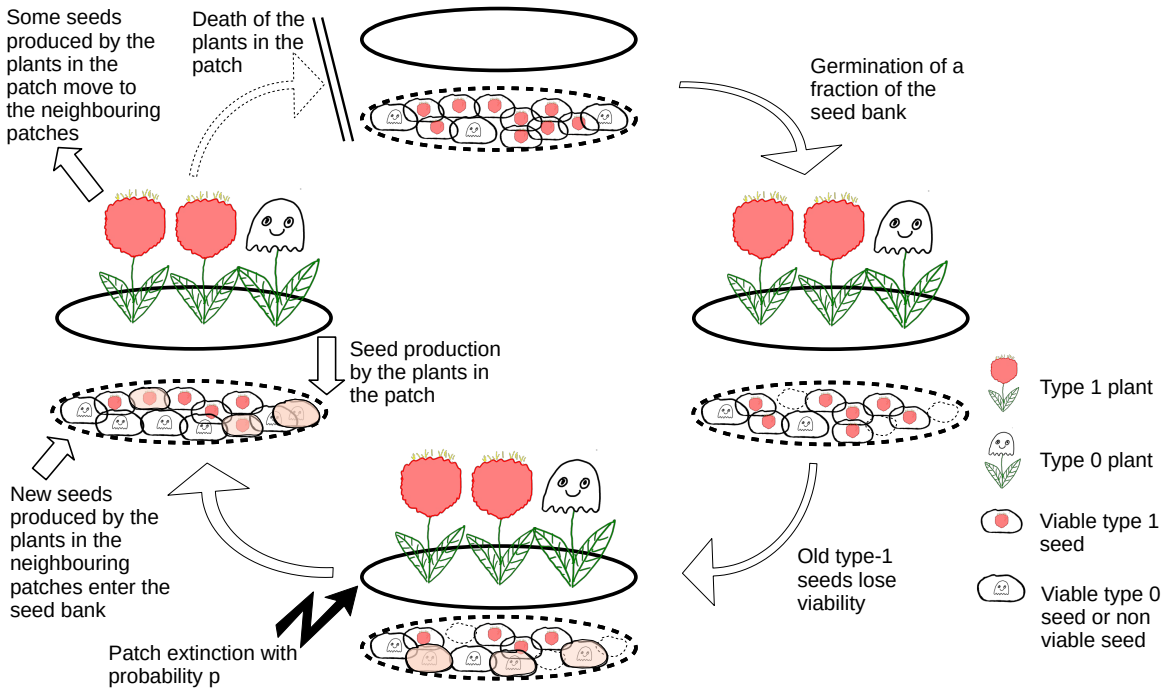


Figure 6.1: Illustration of the intra-patch dynamics of the k-parent WFSB metapopulation process. Here $M = 12$ and $\lfloor gM \rfloor = 3$. The double line in the top line of the figure indicates the starting time of a new generation.

In all that follows, we will refer to :

- M as the *number of seeds per patch*,
- H as the *maximal dormancy duration*,
- g as the *germination probability*,

- c as the *potential colonization probability*,
- p as the *patch extinction probability.*

Moreover, we will say that a patch *goes extinct* during generation n if it is affected by an extinction event during this generation (i.e $\mathrm{Ext}_i = 1$), and that it is *empty* if it does not contain any viable type 1 seeds at the beginning of generation n. Notice that extinction events only affect standing vegetation, and not the seed bank. Therefore, in our terminology, an extinct patch is not necessarily empty. We will also say that the metapopulation *became empty* before generation n if all the patches are empty in generation n, that is, if the subpopulation of real individuals did not persist.

Remark 6.1.2. From a purely mathematical viewpoint, it is also possible to define the model for $k = 1$. However, in this case, real individuals do not have any selective advantage against ghost individuals, and since we assume that the number of real individuals in generation 0 is finite, the metapopulation becomes empty in finite time almost surely.

Remark 6.1.3. Even if the model is defined for $c \in (0, 1/2)$, in practice, since one of the assumptions behind the model is that colonization from patches which are not nearest neighbours is negligible, it is implicitly assumed that c is small. Moreover, notice that if $c > 1/3$, then the potential parents have higher chance of being taken from the patch on the left (or right) than in the focal patch.

Before stating our main results on the model, we give some insight on the reproduction dynamics from a *forwards in time* viewpoint. In order to do so, we first consider a real plant isolated in a patch, surrounded by empty patches. Each empty seed bank compartment in the focal patch chooses the real plant as a potential parent with probability $1-(1-(1-2c)\lfloor gM\rfloor^{-1})^k$, and each empty seed bank compartment in one of the two neighbouring patches chooses it with probability $1-(1-c\lfloor gM\rfloor^{-1})^k$. Therefore, from a forwards in time viewpoint, the number of offspring of the real plant is distributed as the sum of three independent random variables: one binomial random variable with parameters $(\lfloor gM\rfloor, 1-(1-(1-2c)\lfloor gM\rfloor^{-1})^k)$, and two binomial random variables with parameters $(\lfloor gM\rfloor, 1-(1-c\lfloor gM\rfloor^{-1})^k)$.

If $M \to +\infty$ while all other parameters stay constant, the number of offspring is approximately distributed as the sum of three independent Poisson random variables: one with parameter $(1-2c)k$ and two with parameter ck. In particular, the average number of offspring is roughly equal to k, yielding a possible biological interpretation for the parameter k. In this chapter, we focus instead on another parameter regime: $M \to +\infty$ and $k = \lceil M^\alpha \rceil$, $\alpha > 1$. In this regime, the average number of offspring is rather equal to $3\lfloor gM\rfloor$. From a biological viewpoint, this parameter regime corresponds to considering that a plant can potentially produce far more viable seeds than the carrying capacity of a patch.

When several real plants are present, the average number of offspring produced by each plant decreases due to competition. In the case in which competition is the most intense (that is, when the patch and its two neighbours contain only real plants), the number of offspring of each plant in the focal patch is in average equal to 1, and is distributed as the sum of three independent random variables: one binomial random variable with parameters $(\lfloor gM\rfloor, (1-2c)\lfloor gM\rfloor^{-1})$, and two binomial random variables with parameters $(\lfloor gM\rfloor, c\lfloor gM\rfloor^{-1})$. When $M \to +\infty$ and k is fixed, this can be approximated by the sum of a Poisson random variable with parameter $1-2c$ and two Poisson random variables with parameter c. In particular, the random variables do not depend on the parameter k.

Remark 6.1.4. It is possible to generalize the k-parent WFSB metapopulation process by taking the potential parents of a seed in more patches than only neighbouring patches, or by having patches in a two dimensional environment instead of a one dimensional one. If the distance that seeds can travel is bounded, then all the results in this chapter can be extended to the generalized model (though the numerical values for the extinction thresholds will change).

Remark 6.1.5. The idea of sampling several potential parents to model selection can be found in various population genetics models, including variants of the Wright-Fisher model. See e.g [Boe+21a; Boe+21b; CHS19; FP17; GS20; GS18]. Usually, the models comprise both selective reproduction events, during which several potential parents are chosen, and neutral reproduction events, during which only one parent is chosen. Moreover, the mathematical analysis often involves taking selective reproduction events to be rare compared to neutral reproduction events, and to change of time scale to observe them in the limit. In contrast, the model we introduce in this chapter only comprises selective reproduction events, and the questions we aim at answering do not require a change of time scale.

6.1.2 The associated k-parent occupancy process and its limit

BOA process and k-parent occupancy process

The k-parent WFSB metapopulation process can be seen as a multi-colony Wright-Fisher model with selection and seed bank, embedded in a Stochastic Patch Occupancy Model indicating which patches are empty, extinct, or potentially occupied. Indeed, for a patch to contain real seeds, it is not sufficient for it not to go extinct. The viable seeds it contains can only come from 3 patches (the focal patch and its two neighbours), and can only have entered the seed bank during the $H+1$ previous generations. If all these times, the 3 patches were affected by extinction events, then the patch cannot contain viable seeds during the current generation. For instance, if $H=0$, a patch which went extinct along with its two neighbours during the previous generation cannot contain nonexpired type 1 seeds. In the SPOM we define just below, this patch will appear as empty. In other words, the SPOM will encode which patches *cannot* contain type 1 seeds, given the initial condition and the extinction events.

This SPOM is defined on the state space $\mathcal{F}^\infty \times \mathcal{H}^\infty$, with $\mathcal{F}^\infty$ and $\mathcal{H}^\infty$ given by:

$$\mathcal{F}^\infty := \{(O_i)_{i\in\mathbb{Z}} : \forall i \in \mathbb{Z}, O_i \in \{0,1\} \text{ and } \mathrm{Card}\left(\{i \in \mathbb{Z} : O_i = 1\}\right) < +\infty\}$$
$$\text{and } \mathcal{H}^\infty := \{(h_i)_{i\in\mathbb{Z}} : \forall i \in \mathbb{Z}, h_i \in \mathbb{N}\}\,.$$

As for the k-parent WFSB metapopulation process, each patch is associated to a type (0 or 1) and an age, but now they have a different interpretation. Indeed, in the SPOM, a "type 0" patch corresponds to a patch which cannot contain nonexpired type 1 seeds, while a "type 1" patch is a patch which can potentially contain type 1 seeds, the age h_i encoding the last time type 1 seeds could have entered the seed bank.

Definition 6.1.6. *(BOA process) Let $(O,h) \in \mathcal{F}^\infty \times \mathcal{H}^\infty$. The Best Occupancy Achievable process (or BOA process) with parameters (H,p) and with initial conditions (O,h) is the $(\mathcal{F}^\infty \times \mathcal{H}^\infty)$-valued Markov process $(O^{\infty,n}, h^{\infty,n})_{n\in\mathbb{N}}$ defined as follows. First, we set $(O^{\infty,0}, h^{\infty,0}) = (O,h)$. Then, for all $n \in \mathbb{N}$, given $(O^{\infty,n}, h^{\infty,n})$:*

1. *Let $(\mathrm{Ext}_i)_{i\in\mathbb{Z}}$ be i.i.d $\{0,1\}$-valued random variables such that $\mathbb{P}(\mathrm{Ext}_1 = 1) = p$.*
2. *For all $i \in \mathbb{Z}$, if $\mathrm{Ext}_i = 0$ and $O_i^{\infty,n}\mathbb{1}_{\{h_i^{\infty,n}\le H\}} = 1$, then we set*

$$O_{i-1}^{\infty,n+1} = O_i^{\infty,n+1} = O_{i+1}^{\infty,n+1} = 1$$
$$\text{and} \qquad h_{i-1}^{\infty,n+1} = h_i^{\infty,n+1} = h_{i+1}^{\infty,n+1} = 0.$$

   We do nothing during this step if $\mathrm{Ext}_i = 1$ or $O_i^{\infty,n}\mathbb{1}_{\{h_i^{\infty,n}\le H\}} = 0$.

3. *For all $i \in \mathbb{Z}$, if $O_i^{\infty,n+1}$ was not defined during step 2, then we set $O_i^{\infty,n+1} = O_i^{\infty,n}$ and $h_i^{\infty,n+1} = h_i^{\infty,n} + 1$.*

Moreover, we will say that patch $i \in \mathbb{Z}$ is reachable at generation $n \in \mathbb{N}$ if $O_i^{\infty,n} \mathbb{1}_{\{h_i^{\infty,n} \leq H\}} = 1$.

The BOA process represents all the patches which can potentially contain seeds produced by the ones initially present (as given by (O, h)), given the extinction events. In other words, informally, the BOA process keeps track of the patches that are linked to the patches originally containing viable seeds by means of a path of reachable patches. Notice that $O_i^{\infty,n}$ describes the composition of the seed bank, while extinction events affect the standing vegetation. Therefore, an extinction event affecting patch i during the n-th generation does not set the value of $O_i^{\infty,n}$ to 0.

The BOA process is a best-case scenario, in the sense that using the same extinction events to construct the BOA process and the k-parent WFSB metapopulation process, it is possible to couple both processes so that all patches containing seeds in the k-parent WFSB metapopulation process are reachable patches in the BOA process. In order to formalize the coupling property, we introduce a new object associated to our metapopulation process, describing whether the seed bank in each patch contains real seeds, or only ghost seeds.

Definition 6.1.7. *(k-parent occupancy process) Let $(\xi, h) \in \mathcal{F}_M \times \mathcal{H}_M$. The k-parent occupancy process*

$$\left(O^{k,n}, h^{k,n}\right)_{n\in\mathbb{N}} = \left(\left(O_i^{k,n}, h_i^{k,n}\right)_{i\in\mathbb{Z}}\right)_{n\in\mathbb{N}}$$

associated to the k-parent WFSB metapopulation process $(\xi^n, h^n)_{n\in\mathbb{N}}$ with parameters (M, H, g, c, p) and initial conditions (ξ, h) is defined as follows.

First, for all $i \in \mathbb{Z}$, we set

$$O_i^{k,0} := 1 - \prod_{j\in[\![1,M]\!]} (1 - \xi_{i,j}) = \max\{\xi_{i,j} : j \in [\![1, M]\!]\}$$

$$h_i^{k,0} := \begin{cases} \min\{h_{i,j} : j \in [\![1, M]\!] \text{ and } \xi_{i,j} = 1\} & \text{if } O_i^{k,0} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then, for all $n \in \mathbb{N}^*$ and $i \in \mathbb{Z}$, we set

$$O_i^{k,n} := 1 - \prod_{j\in[\![1,M]\!]} \left(1 - \xi_{i,j}^n\right)$$

$$h_i^{k,n} := \begin{cases} \min\{h_{i,j}^n : j \in [\![1, M]\!] \text{ and } \xi_{i,j}^n = 1\} & \text{if } O_i^{k,n} = 1 \\ h_i^{k,n-1} + 1 & \text{otherwise.} \end{cases}$$

In this setting, $O_i^{k,n} = 1$ if and only if at the beginning of generation n, before germination occurs, the patch i contains at least one (potentially expired) seed which was initially of type 1. In this case, $h_i^{k,n}$ is the number of complete generations spent in the seed bank by the youngest of such seeds. Therefore, patch i contains at least one type 1 seed at the beginning of generation n if, and only if it contains seeds which where initially of type 1 (i.e, $O_i^{k,n} = 1$) and the youngest seeds among these ones entered the seed bank at most $H + 1$ generations ago (i.e, $h_i^{k,n} \leq H$, as $h_i^{k,n} = 0$ if the seeds entered the seed bank during the previous generation), that is, if and only if $O_i^{k,n} \mathbb{1}_{\{h_i^{k,n} \leq H\}} = 1$.

Remark 6.1.8. Note that the k-parent occupancy process is also defined on the state space $\mathcal{F}^\infty \times \mathcal{H}^\infty$. However, contrary to the BOA process, the k-parent occupancy process *cannot* be considered as a SPOM, since $(O^{k,n+1}, h^{k,n+1})$ does not depend only on $(O^{k,n}, h^{k,n})$. Therefore, both processes are intrinsically different.

In all that follows, we will say that the BOA process $(O^{\infty,n}, h^{\infty,n})_{n\in\mathbb{N}}$ *associated to* the k-parent WFSB metapopulation process with parameters (M, H, g, c, p) and initial condition (ξ, h) is the BOA

process with parameters (H, p) and initial condition $(O^{k,0}, h^{k,0})$, constructed using the same extinction events as the k-parent WFSB metapopulation process. Under this coupling, the k-parent WFSB metapopulation process and its associated BOA process satisfy the following relation:

$$\forall n \in \mathbb{N}, \forall i \in \mathbb{Z}, O_i^{k,n} \leq O_i^{\infty,n} \text{ and } h_i^{k,n} \geq h_i^{\infty,n}.$$

This result will be proved in Section 6.3.1.

Convergence of the k-parent occupancy process to the BOA process

When M and k are finite, deviations from the BOA process can occur in the following three cases:

1. Type 1 plants are present in a patch, but none of them is chosen as a potential parent.
2. Non-expired type 1 seeds are present in a patch, but none of them germinate.
3. Several type 1 seeds entered the seed bank less than $H+1$ generations ago, but all of them already germinated, and there is no remaining non-expired type 1 seeds in the seed bank.

However, when both $M \to +\infty$ and $k \to +\infty$ in an appropriate way, we can show that the occupancy process converges to the BOA process. For this convergence to occur, two conditions need to be satisfied. First, k needs to grow to $+\infty$ "faster" than M. We will set $k = \lceil M \rceil^\alpha$, with $\alpha > 1$, and hence define a sequence of $\lceil M \rceil^\alpha$-parent WFSB processes. Notice that since the k potential parents of an individual do not have to be necessarily different, it is possible to have $k > 3\lfloor gM \rfloor$ (the number of plants in the focal patch and the two neighbouring patches). Then, we will need the following constraints on the initial conditions of the processes. Let $(O^\infty, h^\infty) \in (\mathcal{F}^\infty \times \mathcal{H}^\infty)$ satisfying

$$\forall i \in \mathbb{Z}, \text{ if } O_i^\infty = 0, \text{ then } h_i^\infty = 0 \tag{6.1.1}$$

encode which patches are initially occupied, and what is the age of the youngest type 1 seeds in each of these patches. By convention, we set $h_i^\infty = 0$ for the patches initially empty. For all $i \in \mathbb{Z}$ such that $O_i^\infty = 1$, let $g_i \in (0, 1]$, which will represent the proportion of youngest type 1 seeds in patch i.

For all $M \geq 2$, let $(\xi^{(M)}, h^{(M)}) \in (\mathcal{F}_M \times \mathcal{H}_M)$ be such that for all $i \in \mathbb{Z}$ and $j \in [\![1, M]\!]$, the following conditions are satisfied.

(A) If $O_i^\infty = 0$, then $\xi_{i,j}^{(M)} = 0$.

(B) If $O_i^\infty = 1$ and $h_{i,j}^{(M)} < h_i^\infty$, then $\xi_{i,j}^{(M)} = 0$.

(C) If $O_i^\infty = 1$, then $M^{-1} \sum_{i=1}^{M} \xi_{i,j}^{(M)} \mathbb{1}_{\{h_{i,j}^{(M)} = h_i^\infty\}} = M^{-1} \lfloor g_i M \rfloor$.

Intuitively, conditions (A), (B) and (C) ensure that the patches initially occupied in all k-parent WFSB metapopulation processes are the same. Conditions (B) and (C) implies that in each patch, the youngest type 1 seeds (if present) have the same age for all processes. Moreover, condition (C) quantifies the proportion of youngest type 1 seeds in the seed bank, and ensures they represent a non-negligible portion of the seed bank, even when $M \to +\infty$. Note that this constraint is on the proportion of the *youngest* type 1 seeds, and not on the proportion of all type 1 seeds.

Theorem 6.1.9. *Let $\alpha > 1$. For all $M \geq 2$, let $\left(O^{(M),n}, h^{(M),n}\right)_{n \in \mathbb{N}}$ be the $\lceil M^\alpha \rceil$-parent occupancy process associated to the $\lceil M^\alpha \rceil$-parent WFSB metapopulation process with parameters*

(M, H, g, c, p) and initial condition $(\xi^{(M)}, h^{(M)})$, and let $\left(O^{(M),\infty,n}, h^{(M),\infty,n}\right)_{n\in\mathbb{N}}$ be the BOA process associated to the same WFSB metapopulation process. Then, for all $N \in \mathbb{N}$,

$$\mathbb{P}\left(\bigcap_{n=0}^{N}\left(\left\{\forall i \in \mathbb{Z}, O_i^{(M),n} = O_i^{(M),\infty,n}\right\} \cap \left\{\forall i \in \mathbb{Z}, h_i^{(M),n} = h_i^{(M),\infty,n}\right\}\right)\right) \xrightarrow[M\to+\infty]{} 1.$$

One of the biological interpretations of this result is that in the limit considered, the metapopulation dynamics is well approximated by the BOA process. Moreover, this theorem bridges the gap between individual-based metapopulation models and SPOMs, in the sense that the BOA process is the limit of the k-parent WFSB metapopulation process under a suitable scaling.

Sketch of the proof. The proof is structured as follows. First, we show that if the number of occupied patches is finite initially (as assumed in Condition (A)), then it remains finite. The proof relies on the observation that colonization is only possible towards or from neighbouring patches.

Then, we show that when $k = \lceil M^\alpha \rceil$ is large enough, in each patch, the k potential parents chosen to refill any seed bank compartment in the patch span all the $3\lfloor gM \rfloor$ possible potential parents with high probability. In this case, the $\lfloor gM \rfloor$ new seeds entering the seed bank of the patch are all of the same type. Therefore, we can distinguish two periods:

- A transition period corresponding to the $H + 1$ first generations, during which some seeds initially present are still in the seed bank and viable, implying that some viable seeds of the same age (and in the same patch) are potentially of different types;
- What we will call the "post-transition period", corresponding to the other generations, during which all the seeds initially present and still in the seed bank are no longer viable, and during which all viable seeds of the same age and in the same patch are of the same type (with an high probability).

We first focus on the proportion of viable seeds of a given age in the seed bank after the transition period. We show that for all $0 \leq h \leq H$, the proportion of age h seeds (which are generally all of the same type) is roughly equal to $g(1-g)^h$, and that approximately $g^2(1-g)^h M$ (> 1 for M large enough) such seeds germinate during the next generation. Notice that it is what we would obtain if each seed germinated independently from others and with probability g. In particular, if type 1 seeds enter the seed bank of a patch during generation n, then with high probability at least one of them will germinate during each of the generations $n + 1$, ..., $n + H + 1$, and produce new seeds if the patch is not affected by an extinction event, as in the BOA process.

We conclude by considering the transition period. At the beginning of generation $n \in [\![0, H]\!]$, in each non-empty patch:

- Either no new type 1 seeds have already entered the seed bank. By conditions (B) and (C), the proportion of youngest type 1 seeds is then roughly equal to $g_i(1-g)^n$, and approximately $gg_i(1-g)^n M$ (> 1 for M large enough) such seeds germinate during the next generation.
- Either new type 1 seeds have already entered the seed bank. We are then in the same situation as during the post-transition period.

Critical patch extinction probability

Using the coupling with the BOA process, we will also show the existence of a critical patch extinction probability $p_{crit}(H)$ depending only on H such that for all $p > p_{crit}(H)$, the metapopulation will almost surely become empty in finite time no matter the values of M, g, c or k.

Theorem 6.1.10. *For all $H \in \mathbb{N}$, there exists $p_{crit}(H) \in (0,1)$ such that for all $M \in \mathbb{N}^*$, $k \in \mathbb{N}\backslash\{0,1\}$, $g \in (0,1)$ and $c \in (0,1/2)$, for all $(\xi, h) \in \mathcal{F}_M \times \mathcal{H}_M$ and $p > p_{crit}(H)$, if $(O^{k,n}, h^{k,n})_{n\in\mathbb{N}}$ is the k-parent occupancy process associated to the k-parent WFSB metapopulation process with parameters (M, H, g, c, p) and initial condition (ξ, h), then*

$$\lim_{n\to+\infty} \mathbb{P}\left(\forall i \in \mathbb{Z}, O_i^{k,n} \mathbb{1}_{\{h_i^{k,n} \leq H\}} = 0\right) = 1.$$

The proof of this result, which can be found in Section 6.3, relies on the coupling between the k-parent WFSB metapopulation process and the BOA process, together with appropriate results in percolation theory.

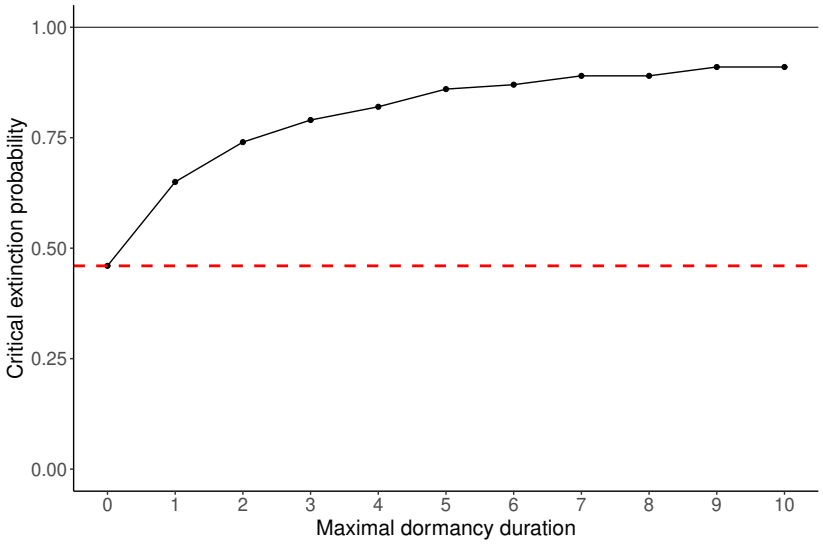


Figure 6.2: Approximate value of the critical extinction probability $p_{crit}(H)$ as a function of the maximal dormancy duration H. The red dashed line indicates $p_{crit}(0)$, or in other words, the extinction probability above which (real) plants persistence without a seed bank is not possible. The continuous line indicates $p_{crit}(\infty) = 1$ (see Remark 6.1.11). See the Appendix for details on the method used to compute $p_{crit}(H)$.

Sketch of the proof. The main idea behind the proof is the following. First, we use the coupling between the k-parent WFSB metapopulation process and the BOA process introduced earlier. This coupling ensures that each occupied patch in the k-parent WFSB metapopulation process is reachable in the BOA process. We can then focus on the study of the simpler BOA process, since the fact that no sites are reachable in the BOA process will then guarantee that the k-parent WFSB metapopulation process has become empty. Using results from percolation theory, we obtain the existence of a critical extinction probability $p_{crit}(H)$ for the BOA process, such that:

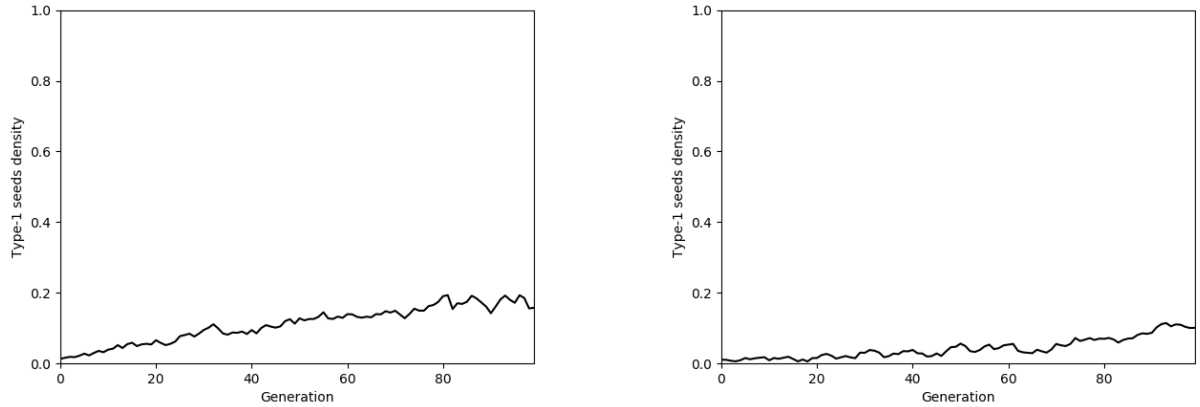
- For all $p > p_{crit}(H)$, the BOA process goes extinct in finite time almost surely ;
- For all $p < p_{crit}(H)$, the probability that the BOA process goes extinct in finite time is (strictly) less than 1.

We conclude by using the fact that the k-parent WFSB metapopulation process is embedded in the BOA process.

The biological interpretation of this theorem is the following. For each maximal dormancy duration $H \in \mathbb{N}$, there exists a critical extinction probability $p_{crit}(H)$ above which any metapopulation

evolving according to a k-parent WFSB metapopulation process of maximal dormancy duration H will almost surely go extinct in finite time, no matter how quickly plants can invade a patch initially empty (which is quantified by k and to a lesser extent c). In particular, no metapopulation without a seed bank can persist if the patch extinction probability is above $p_{crit}(0)$. $p_{crit}(H)$ is increasing with H, so the ability to form a seed bank can potentially allow population persistence and expansion in highly disturbed fragmented environments. See Figure 6.2 for approximate values for $p_{crit}(H)$, computed using the method presented in the Appendix. Numerical simulations show the existence of parameter sets (M, H, g, c, p) with $H > 0$ and $p > p_{crit}(0)$ for which population persistence is indeed possible (see Figure 6.3). Since the k-parent occupancy process converges to the BOA process, which does not go extinct with positive probability if $p < p_{crit}(H)$, the critical extinction probability $p_{crit}(H)$ we obtain is optimal, in the sense that it is not possible to obtain a better upper bound on the critical extinction probability for the k-parent WFSB metapopulation process which depends only on H, and not also on one of the other parameters.

Remark 6.1.11. In the limiting case $H = \infty$, seeds never expire, and the number of reachable patches in the BOA process cannot decrease, implying that $p_{crit}(\infty) = 1$.



(a) p = 0.5 and H = 1

(b) p = 0.7 and H = 3

Figure 6.3: Plant metapopulation expansion for extinction probabilities $p_{crit}(0) < p < p_{crit}(H)$, and for a maximal dormancy duration $H \neq 0$. The values taken by the other parameters are $M = 100$, $g = 0.5$, $c = 0.05$ and $k = 25$. Initially, 5 consecutive patches contained $gM = 50$ type 1 seeds, and all the other seed bank compartments were empty. Since only the first 100 generations were considered, the simulation was performed on a torus of 200 patches, and the density of type 1 seeds was computed over these 200 patches.

6.2 Proof of the convergence of the k-parent occupancy process to the BOA process

The goal of this section is to show that the k-parent occupancy process converges to the BOA process in the sense of Theorem 6.1.9, that is, when both $M \to +\infty$ and $k \to +\infty$, but with k increasing "faster" than M. In order to do so, we first focus on the post-transition period, and take an initial condition allowing us to skip the transition period. Then, we explain how to adapt the proof to take into account more general initial conditions.

First, we set some notation. We use the notation from Theorem 6.1.9 throughout this section. For all $M \geq 2$ and $H \in \mathbb{N}$, let $(\xi^{(M),n}, h^{(M),n})_{n\in\mathbb{N}}$ be the $\lceil M^\alpha \rceil$-parent WFSB metapopu-

lation process with parameters (M, H, g, c, p) associated to the $\lceil M^\alpha \rceil$-parent occupancy process $(O^{(M),n}, h^{(M),n})_{n\in\mathbb{N}}$. For every $n \in \mathbb{N}$, $i \in \mathbb{Z}$ and $h \in \mathbb{N}$, let

$$E_{i,h}^{(M),n} := \{j \in [\![1, M]\!] : h_{i,j}^{(M),n} = h\}$$

be the set of all seed bank compartments in patch i containing seeds of age h at the beginning of generation n, and let $G_i^{(M),n}$ be the set of all seed bank compartments in patch i which contain the seeds germinating at the beginning of generation n.

Recalling that $(O^{(M),\infty,n}, h^{(M),\infty,n})_{n\in\mathbb{N}}$ is the BOA process associated to $(\xi^{(M),n}, h^{(M),n})_{n\in\mathbb{N}}$, we introduce the event

$$\mathrm{DiffBOA}_i^{(M),n} := \left\{ O_i^{(M),n} \neq O_i^{(M),\infty,n} \text{ or } h_i^{(M),n} \neq h_i^{(M),\infty,n} \right\}.$$

Given the initial condition for the associated BOA process, we already know that for $n = 0$ the BOA process associated to $(\xi^{(M),n}, h^{(M),n})_{n\in\mathbb{N}}$ is equal to the occupancy process associated to the same process. Therefore,

$$\mathbb{P}\left(\bigcap_{i\in\mathbb{Z}} \overline{\mathrm{DiffBOA}_i}^{(M),0} \right) = 1,$$

where the bar denotes the complementary event. Moreover, as only neighbouring sites can send colonizing seeds, if type 1 seeds are initially in a finite number of patches, then it is also the case after any arbitrary finite duration. More specifically, if we set

$$i_{min}^0 := \min\{i \in \mathbb{Z} : O_i^\infty = 1\} \text{ and } i_{max}^0 := \max\{i \in \mathbb{Z} : O_i^\infty = 1\},$$

and if for all $n \in \mathbb{N}$, we set

$$i_{min}^{n+1} := i_{min}^n - 1 \text{ and } i_{max}^{n+1} := i_{max}^n + 1,$$

then the only patches which can potentially contain type 1 seeds after n generations are the patches with index $i \in [\![i_{min}^n, i_{max}^n]\!]$. In other words, for all $M \geq 2$, $n \in \mathbb{N}$ and $i \in \mathbb{Z}\backslash[\![i_{min}^n, i_{max}^n]\!]$, we have $O_i^{(M),n} = O_i^{(M),\infty,n} = 0$ and $h_i^{(M),n} = h_i^{\infty,n}$, so that $\overline{\mathrm{DiffBOA}}_i^{(M),n}$ holds a.s. Therefore, for all $N \in \mathbb{N}$,

$$\mathbb{P}\left(\bigcup_{n=0}^{N} \bigcup_{i\in\mathbb{Z}} \mathrm{DiffBOA}_i^{(M),n} \right) = \mathbb{P}\left(\bigcup_{n=1}^{N} \bigcup_{i\in[\![i_{min}^n, i_{max}^n]\!]} \mathrm{DiffBOA}_i^{(M),n} \right). \tag{6.2.1}$$

In all that follows, in order to ease notation, for all $n \in \mathbb{N}$, we set $I_n = [\![i_{min}^n, i_{max}^n]\!]$

Let $M \geq 2$ such that $\lfloor gM \rfloor > 1$, and $N \geq 1$. We recall that deviations from the BOA process can occur in the following cases.

1. Type 1 plants are present in a patch, but are never chosen as potential parents.
2. The seed bank contains non-expired type 1 seeds, but none of them germinate during the generation we consider.
3. Some type 1 seeds did enter the seed bank less than $H + 1$ generations ago, but all of them already germinated, and the seed bank is now empty.

Let $(a_n)_{n\in\mathbb{N}}$ be the sequence defined by

$$a_0 = 0 \text{ and } \forall n \in \mathbb{N}, a_{n+1} = 3a_n + 1,$$

and for all $0 < \epsilon < 1$, let

$$W_h^{(M),\epsilon} := \left[\lfloor gM \rfloor \left(\left(1 - \frac{\lfloor gM \rfloor}{M} \right)^h - \epsilon a_h \right), \lfloor gM \rfloor \left(\left(1 - \frac{\lfloor gM \rfloor}{M} \right)^h + \epsilon a_h \right) \right].$$

If the initial condition $(\xi^{(M),0}, h^{(M),0})$ satisfies the extra hypothesis:

- (IC1) All the viable seeds of the same age and in the same patch are of the same type, or in formula,
$$\forall i \in \mathbb{Z}, \forall j_1, j_2 \in [\![1, M]\!], h_{i,j_1}^{(M),0} = h_{i,j_2}^{(M),0} \leq H \Longrightarrow \xi_{i,j_1}^{(M),0} = \xi_{i,j_2}^{(M),0};$$

- (IC2) For all $h \in [\![0, H]\!]$, the proportion of age h seeds is roughly equal to $g(1-g)^h$, in the sense that
$$\mathrm{Card}(E_{i,h}^{(M),0}) \in W_h^{(M),\epsilon}$$

for ϵ appropriately chosen, then the three deviation cases are covered by the following more general cases:

- (D1) There exists $n \in \mathbb{N}$ and $i \in I_{n+1}$ such that one of the plants in patches $\{i-1, i, i+1\}$ is not chosen as a potential parent by at least one seed bank compartment in patch i during generation n.

- (D2) There exists $n \in \mathbb{N}$ and $i \in I_n$ such that $h_i^{(M),n} \leq H$ but no seed of age $h_i^{(M),n}$ germinates during generation n.

- (D3) There exists $n \in \mathbb{N}$ and $i \in I_n$ such that $h_i^{(M),n} \leq H$ but $\mathrm{Card}\left(E_{i,h_i^{(M),n}}^{(M),n} \right) \notin W_{h_i^{(M),n}}^{(M),\epsilon}$.

Indeed, initially, all the viable seeds of the same age have the same type, and this stays true until (D1) occurs. If (D1) did not occur yet before generation n, then for all $i \in I_n$, all seeds of age $h_i^{(M),n}$ are of the same type. They are not necessarily type 1 seeds, but if the patch is not empty, then they are all type 1 seeds. Therefore, if the patch is not empty and if ϵ is such that

$$\lfloor gM \rfloor \left(\left(1 - \lfloor gM \rfloor M^{-1} \right)^{h_i^{(M),n}} - \epsilon a_{h_i^{(M),n}} \right) > 1,$$

then

$$\mathrm{Card}\left(E_{i,h_i^{(M),n}}^{(M),n} \right) \in W_{h_i^{(M),n}}^{(M),\epsilon}$$

implies that the patch contains type 1 seeds which entered less than $H+1$ generations ago and did not germinate yet, and

$$\mathrm{Card}\left(E_{i,h_i^{(M),n}}^{(M),n} \cap G_i^{(M),n} \right) > 0$$

implies that at least one type 1 seed germinates during generation n.

Formally, let $\epsilon > 0$ be such that for all $M \geq 2$ such that $\left(1 - \lfloor gM \rfloor M^{-1} \right)^H - \lfloor gM \rfloor^{-1} > 0$,

$$\epsilon < a_H^{-1} \left(\left(1 - \lfloor gM \rfloor M^{-1} \right)^H - \lfloor gM \rfloor^{-1} \right).$$

For all $n \in \mathbb{N}$ and $i \in \mathbb{Z}$, we define the following events corresponding respectively to (D1), (D2) and (D3):

$$\begin{aligned}
\mathrm{Par}_i^{(M),n} &:= \{\text{There exists a plant in patches } \{i-1,i,i+1\} \text{ which is not chosen as a} \\
&\qquad \text{potential parent by at least one seed bank compartment in patch i during} \\
&\qquad \text{generation n}\}\,, \\
\mathrm{ErrG}_i^{(M),n} &:= \left\{ h_i^{(M),n} \leq H \text{ and } \mathrm{Card}\left(E^{(M),n}_{i,h_i^{(M),n}} \cap G_i^{(M),n} \right) = 0 \right\}, \\
\text{and} \quad \mathrm{ErrF}_i^{(M),n} &:= \left\{ h_i^{(M),n} \leq H \text{ and } \mathrm{Card}\left(E^{(M),n}_{i,h_i^{(M),n}} \right) \notin W^{(M),\epsilon}_{h_i^{(M),n}} \right\}.
\end{aligned}$$

We assume that the initial condition $(\xi^{(M),0}, h^{(M),0})$ satisfies conditions (IC1) and (IC2). Then,

$$\mathbb{P}\left(\bigcup_{n=1}^{N} \bigcup_{i \in I_n} \mathrm{DiffBOA}_i^n \right) \leq \mathbb{P}\left(\bigcup_{n=0}^{N-1} \bigcup_{i \in I_{n+1}} \mathrm{Par}_i^{(M),n} \cup \mathrm{ErrG}_i^{(M),n} \cup \mathrm{ErrF}_i^{(M),n+1} \right). \tag{6.2.2}$$

Remark 6.2.1. If $(\xi^{(M),0}, h^{(M),0})$ does not satisfy condition (IC1) or (IC2), then we show at the end of the section that after a transition period of $H+1$ generations, $(\xi^{(M),H+1}, h^{(M),H+1})$ satisfies conditions (IC1) and (IC2) with high probability. We can then restart the process from $(\xi^{(M),H+1}, h^{(M),H+1})$ and conclude.

In order to shorten the notation, let $\mathrm{ErrGF}_i^{(M),n}$ be the event defined as

$$\mathrm{ErrGF}_i^{(M),n} := \mathrm{ErrG}_i^{(M),n} \cup \mathrm{ErrF}_i^{(M),n+1}.$$

Since the choice of the potential parents does not depend on their types and is independent from one patch (or generation) to another, we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{n=0}^{N-1} \bigcup_{i \in I_{n+1}} \mathrm{Par}_i^{(M),n} \cup \mathrm{ErrGF}_i^{(M),n} \right) \\
&\leq \sum_{n=0}^{N-1} \sum_{i \in I_{n+1}} \mathbb{P}\left(\mathrm{Par}_i^{(M),n} \right) + \mathbb{P}\left(\bigcup_{n=0}^{N-1} \bigcup_{i \in I_{n+1}} \mathrm{ErrGF}_i^{(M),n} \,\middle|\, \left(\overline{\mathrm{Par}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq N-1 \\ i \in I_{n'+1}}} \right) && (6.2.3) \\
&\leq \sum_{n=0}^{N-1} \left(i^0_{max} - i^0_{min} + 1 + 2(n+1) \right) \mathbb{P}\left(\mathrm{Par}^{(M),0}_{i^0_{min}} \right) + \mathbb{P}\left(\bigcup_{i \in I_1} \mathrm{ErrGF}_i^{(M),0} \,\middle|\, \left(\overline{\mathrm{Par}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq N-1 \\ i \in I_{n'+1}}} \right) \\
&\qquad + \sum_{n=1}^{N-1} \mathbb{P}\left(\bigcup_{i \in I_{n+1}} \mathrm{ErrGF}_i^{(M),n} \,\middle|\, \left(\overline{\mathrm{Par}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq N-1 \\ i \in I_{n'+1}}} \cap \left(\overline{\mathrm{ErrGF}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq n-1 \\ i \in I_{n'+1}}} \right) \\
&\leq \left[N\left(i^0_{max} - i^0_{min} + 1 \right) + N(N+1) \right] \mathbb{P}\left(\mathrm{Par}^{(M),0}_{i^0_{min}} \right) && (6.2.4) \\
&\qquad + \mathbb{P}\left(\bigcup_{i \in I_1} \mathrm{ErrGF}_i^{(M),0} \,\middle|\, \left(\overline{\mathrm{Par}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq N-1 \\ i \in I_{n'+1}}} \right) && (6.2.5) \\
&\qquad + \sum_{n=1}^{N} \mathbb{P}\left(\bigcup_{i \in I_{n+1}} \mathrm{ErrGF}_i^{(M),n} \,\middle|\, \left(\overline{\mathrm{Par}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq N-1 \\ i \in I_{n'+1}}} \cap \left(\overline{\mathrm{ErrGF}}_i^{(M),n'} \right)_{\substack{0 \leq n' \leq n-1 \\ i \in I_{n'+1}}} \right). && (6.2.6)
\end{aligned}$$

Bounding the quantity in (6.2.4) from above is straightforward, and is carried out in Section 6.2.1. The main obstacle to the study of the two other terms stems from the fact that all the events considered are linked together by $(h^{(M),n})_{n\in\mathbb{N}}$. However, we can circumvent this problem by working conditionally on $h^{(M),n}$. Indeed, given $h_i^{(M),n}$, since $h_i^{(M),n+1}$ can only be equal to 0 or $h_i^{(M),n}+1$,

$$\begin{aligned}\mathrm{ErrF}_i^{n+1} &:= \left\{\mathrm{Card}\left(E^{(M),n+1}_{i,h_i^{(M),n+1}}\right) \notin W^{(M),\epsilon}_{h_i^{(M),n+1}} \text{ and } h_i^{(M),n+1} \leq H\right\}\\ &\subseteq \left\{\mathrm{Card}\left(E^{(M),n+1}_{i,0}\right) \notin W_0^{(M),\epsilon} \text{ and } 0 \leq H\right\}\\ &\quad \cup \left\{\mathrm{Card}\left(E^{(M),n+1}_{i,h_i^{(M),n}+1}\right) \notin W^{(M),\epsilon}_{h_i^{(M),n}+1} \text{ and } h_i^{(M),n}+1 \leq H\right\}\\ &= \left\{\mathrm{Card}\left(E^{(M),n+1}_{i,h_i^{(M),n}+1}\right) \notin W^{(M),\epsilon}_{h_i^{(M),n}+1} \text{ and } h_i^{(M),n}+1 \leq H\right\}\end{aligned}$$

since $W_0^{(M),\epsilon} = \{\lfloor gM \rfloor\} = \{\mathrm{Card}(E_{i,0}^{(M),n+1})\}$. Therefore,

$$\mathrm{ErrGF}_i^n \subseteq \{h_i^{(M),n} \leq H\} \cap \left(\left\{\mathrm{Card}\left(E^{(M),n}_{i,h_i^{(M),n}} \cap G_i^{(M),n}\right) = 0\right\} \cup \left\{\mathrm{Card}\left(E^{(M),n+1}_{i,h_i^{(M),n}+1}\right) \notin W^{(M),\epsilon}_{h_i^{(M),n}+1}\right\}\right). \tag{6.2.7}$$

In order to use (6.2.7), in Section 6.2.2, we establish an upper bound on

$$\mathbb{P}\left(\left\{\mathrm{Card}\left(E_{i,h}^{(M),n} \cap G_i^{(M),n}\right) = 0\right\} \cup \left\{\mathrm{Card}\left(E_{i,h+1}^{(M),n+1}\right) \notin W_{h+1}^{(M),\epsilon}\right\} \middle| \left\{\mathrm{Card}\left(E_{i,h}^{(M),n}\right) \in W_h^{(M),\epsilon}\right\}\right) \tag{6.2.8}$$

for all $h \in [\![0, H]\!]$, which does not depend on the value of h. In Section 6.2.3, we use it, combined with the upper bound on (6.2.4) from Section 6.2.1, in order to complete the proof of Theorem 6.1.9.

6.2.1 Upper bound on $\mathbb{P}(\mathrm{Par}_i^{(M),n})$

We set $c^* = \min(c, 1-2c)$. The goal of this section is to show the following lemma.

Lemma 6.2.2. *For all $M \geq 2$, $n \in \mathbb{N}$ and $i \in \mathbb{Z}$,*

$$\mathbb{P}\left(\mathrm{Par}_i^{(M),n}\right) \leq 3g^2M^2 \exp\left(M^\alpha \ln\left(1 - \frac{c^*}{gM}\right)\right).$$

A direct consequence of this lemma is the fact that since $\alpha > 1$,

$$\mathbb{P}\left(\mathrm{Par}_i^{(M),n}\right) \xrightarrow[M\to+\infty]{} 0.$$

Proof. Assume that $c^* = c$. Let $\widetilde{\mathrm{Par}}_i^{(M),n}$ be the event: {"The first seed which germinated in patch $i+1$ was not chosen as a potential parent by the first seed bank compartment in patch i to be refilled"}. Then,

$$\mathbb{P}\left(\mathrm{Par}_i^{(M),n}\right) \leq 3\lfloor gM \rfloor^2 \mathbb{P}\left(\widetilde{\mathrm{Par}}_i^{(M),n}\right).$$

Indeed, $\mathrm{Par}_i^{(M),n}$ is the event {"at least one plant in one of the patches $\{i-1, i, i+1\}$ is not chosen as a potential parent in order to refill at least one seed bank compartment in patch i"}. There exists $3\lfloor gM \rfloor^2$ pairs "plant not chosen in patch $i-1$, i or $i+1$/seed bank compartment in patch i", and as $c^* = c$, plants in patches $i-1$ and $i+1$ have less chances of being chosen as potential parents than plants in patch i.

Then, each one of the $\lceil M^\alpha \rceil$ potential parents chosen to refill the first seed bank compartment in patch i is *not* the first plant of patch $i+1$ with probability

$$1 - \frac{c^*}{\lfloor gM \rfloor} \leq 1 - \frac{c^*}{gM}.$$

Hence,

$$\begin{aligned} \mathbb{P}\left(\widetilde{\mathrm{Par}}_i^{(M),n}\right) &\leq \left(1 - \frac{c^*}{gM}\right)^{\lceil M^\alpha \rceil} \\ &\leq \left(1 - \frac{c^*}{gM}\right)^{M^\alpha} \\ \text{and} \qquad \mathbb{P}\left(\mathrm{Par}_i^{(M),n}\right) &\leq 3g^2M^2 \exp\left(M^\alpha \ln\left(1 - \frac{c^*}{gM}\right)\right). \end{aligned}$$

If $c^* \neq c$, then we can directly adapt this proof defining instead the event $\widetilde{\mathrm{Par}}_i^{(M),n}$ as the event {"The first seed which germinated in the patch i (instead of the patch $i+1$) was not chosen as a potential parent by the first seed bank compartment in patch i to be refilled"}. □

6.2.2 Upper bound on (6.2.8)

In this section, we show an upper bound on

$$\mathbb{P}\left(\left\{\mathrm{Card}\left(E_{i,h}^{(M),n} \cap G_i^{(M),n}\right) = 0\right\} \cup \left\{\mathrm{Card}\left(E_{i,h+1}^{(M),n+1}\right) \notin W_{h+1}^{(M),\epsilon}\right\} \middle| \left\{\mathrm{Card}\left(E_{i,h}^{(M),n}\right) \in W_h^{(M),\epsilon}\right\}\right)$$

for all $h \in [\![0, H]\!]$. In other words, we study the probability that if the number of age h seeds in patch i at the beginning of generation n is roughly equal to $g(1-g)^h M$:

- either no age h seeds germinate during generation n
- or the number of remaining age h seeds (which become age $h+1$ seeds in the subsequent generation) is significantly different from $g(1-g)^{h+1}M$.

In order to do so, we first observe that if $E \subseteq [\![1, M]\!]$ is a non-empty strict subset of $[\![1, M]\!]$, then for all $n \in \mathbb{N}$ and $i \in \mathbb{Z}$, $\mathrm{Card}(E \cap G_i^{(M),n})$ follows an hypergeometric law. Using the tail inequalities (10) and (14) from [Ska13] yields the following lemma.

Lemma 6.2.3. *Let $M \geq 2$ be such that $\left(1 - \lfloor gM \rfloor M^{-1}\right)^H - \lfloor gM \rfloor^{-1} > 0$. Let $E \subseteq [\![1, M]\!]$ be a non-empty strict subset of $[\![1, M]\!]$. Then, for all $n \in \mathbb{N}$, $i \in \mathbb{Z}$ and $0 < \epsilon < 1$,*

$$\begin{aligned} &\mathbb{P}\left(\mathrm{Card}(E \cap G_i^{(M),n}) \geq \mathrm{Card}(E)(1+\epsilon)\frac{\lfloor gM \rfloor}{M}\right) \leq e^{-2\epsilon^2 \mathrm{Card}(E)^2 \lfloor gM \rfloor M^{-2}} \\ \textit{and} \quad &\mathbb{P}\left(\mathrm{Card}(E \cap G_i^{(M),n}) \leq \mathrm{Card}(E)(1-\epsilon)\frac{\lfloor gM \rfloor}{M}\right) \leq e^{-2\epsilon^2 \mathrm{Card}(E)^2 \lfloor gM \rfloor M^{-2}} \end{aligned}$$

Using this lemma, we can obtain the following upper bound.

Lemma 6.2.4. *For all $M \geq 2$ such that $\left(1 - \lfloor gM \rfloor M^{-1}\right)^H - \lfloor gM \rfloor^{-1} > 0$, for all $n \in \mathbb{N}$, $i \in \mathbb{Z}$ and $h \in [\![0, H]\!]$,*

$$\begin{aligned} &\mathbb{P}\left(\left\{\mathrm{Card}\left(E_{i,h}^{(M),n} \cap G_i^{(M),n}\right) = 0\right\} \cup \left\{\mathrm{Card}\left(E_{i,h+1}^{(M),n+1}\right) \notin W_{h+1}^{(M),\epsilon}\right\} \middle| \left\{\mathrm{Card}\left(E_{i,h}^{(M),n}\right) \in W_h^{(M),\epsilon}\right\}\right) \\ &\leq 2e^{-2\epsilon^2 \lfloor gM \rfloor^3 ((1 - \lfloor gM \rfloor M^{-1})^H - \epsilon a_H)^2 M^{-2}}. \end{aligned}$$

Proof. Let $h \in [\![0, H]\!]$, $n \in \mathbb{N}$ and $i \in \mathbb{Z}$. Assume that $\mathrm{Card}(E_{i,h}^{(M),n}) \in W_h^{(M),\epsilon}$. We make the following observation.

1. If less than $\lfloor gM \rfloor M^{-1}(1+\epsilon)\mathrm{Card}(E_{i,h}^{(M),n})$ age h seeds germinate during generation n, the number of remaining age h seeds is bounded from below by

$$\begin{aligned}
&\lfloor gM \rfloor \left[\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h - \epsilon a_h - \frac{\lfloor gM \rfloor}{M}(1+\epsilon)\left(\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h + \epsilon a_h\right)\right] \\
&= \lfloor gM \rfloor \left[\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h \left(1 - \frac{\lfloor gM \rfloor}{M}\right) - \epsilon a_h - \frac{\lfloor gM \rfloor}{M}\epsilon a_h - \frac{\lfloor gM \rfloor}{M}\epsilon \left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h - \frac{\lfloor gM \rfloor}{M}\epsilon^2 a_h\right] \\
&\geq \lfloor gM \rfloor \left[\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^{h+1} - \epsilon(3a_h + 1)\right] \\
&= \lfloor gM \rfloor \left(\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^{h+1} - \epsilon a_{h+1}\right)
\end{aligned}$$

   by definition of $(a_h)_{h\geq 0}$. By Lemma 6.2.3, this event happens with probability bounded from below by

$$1 - e^{-2\epsilon^2 \lfloor gM \rfloor^3 ((1 - \lfloor gM \rfloor M^{-1})^h - \epsilon a_h)^2 M^{-2}} \geq 1 - e^{-2\epsilon^2 \lfloor gM \rfloor^3 ((1 - \lfloor gM \rfloor M^{-1})^H - \epsilon a_H)^2 M^{-2}}.$$

2. If more than $\lfloor gM \rfloor M^{-1}(1-\epsilon)\mathrm{Card}(E_{i,h}^{(M),n})$ age h seeds germinate, the number of remaining age h seeds is bounded from above by

$$\begin{aligned}
&\lfloor gM \rfloor \left[\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h + \epsilon a_h - \frac{\lfloor gM \rfloor}{M}(1-\epsilon)\left(\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h - \epsilon a_h\right)\right] \\
&= \lfloor gM \rfloor \left[\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h \left(1 - \frac{\lfloor gM \rfloor}{M}\right) + \epsilon a_h + \frac{\lfloor gM \rfloor}{M}\epsilon a_h + \frac{\lfloor gM \rfloor}{M}\epsilon \left(1 - \frac{\lfloor gM \rfloor}{M}\right)^h - \frac{\lfloor gM \rfloor}{M}\epsilon^2 a_h\right] \\
&\leq \lfloor gM \rfloor \left[\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^{h+1} + \epsilon(3a_h + 1)\right] \\
&= \lfloor gM \rfloor \left(\left(1 - \frac{\lfloor gM \rfloor}{M}\right)^{h+1} + \epsilon a_{h+1}\right).
\end{aligned}$$

   Again by Lemma 6.2.3, this event happens with probability bounded from below by

$$1 - e^{-2\epsilon^2 \lfloor gM \rfloor^3 ((1 - \lfloor gM \rfloor M^{-1})^H - \epsilon a_H)^2 M^{-2}}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left\{ \mathrm{Card}\left(E_{i,h}^{(M),n} \cap G_i^{(M),n}\right) = 0 \right\} \cup \left\{ \mathrm{Card}\left(E_{i,h+1}^{(M),n+1}\right) \notin W_{h+1}^{(M),\epsilon} \right\} \middle| \left\{ \mathrm{Card}\left(E_{i,h}^{(M),n}\right) \in W_h^{(M),\epsilon} \right\} \right) \\
&\leq \mathbb{P}\left(\mathrm{Card}\left(E_{i,h}^{(M),n} \cap G_i^{(M),n}\right) < \frac{\lfloor gM \rfloor}{M}(1-\epsilon)\mathrm{Card}(E_{i,h}^{(M),n}) \middle| \left\{ \mathrm{Card}\left(E_{i,h}^{(M),n}\right) \in W_h^{(M),\epsilon} \right\} \right) \\
&\quad + \mathbb{P}\left(\mathrm{Card}\left(E_{i,h}^{(M),n} \cap G_i^{(M),n}\right) > \frac{\lfloor gM \rfloor}{M}(1+\epsilon)\mathrm{Card}(E_{i,h}^{(M),n}) \middle| \left\{ \mathrm{Card}\left(E_{i,h}^{(M),n}\right) \in W_h^{(M),\epsilon} \right\} \right) \\
&\leq 2e^{-2\epsilon^2 \lfloor gM \rfloor^3 ((1 - \lfloor gM \rfloor M^{-1})^H - \epsilon a_H)^2 M^{-2}},
\end{aligned}$$

and we can conclude. □

6.2.3 Proof of Theorem 6.1.9

In order to show Theorem 6.1.9, we first condition on $h^{(M),n}$ in Eq. (6.2.6). In order to do so, we introduce the following notation. For all $n < N \in \mathbb{N}$, let $\mathrm{Cond}^{(M),n,N}$ be the event defined by

$$\mathrm{Cond}^{(M),n,N} := \left(\overline{\mathrm{Par}}_i^{(M),n'}\right)_{\substack{0\leq n'\leq N-1 \\ i\in I_{n'+1}}} \cap \left(\overline{\mathrm{ErrGF}}_i^{(M),n'}\right)_{\substack{0\leq n'\leq n-1 \\ i\in I_{n'+1}}}. \tag{6.2.9}$$

Moreover, let $\mathcal{V}^{(M),n,N}$ be the set of all possible values for $(h_i^{(M),n})_{i\in I_{n+1}}$ given the initial condition and $\mathrm{Cond}^{(M),n,N}$. That is,

$$\mathcal{V}^{(M),n} := \left\{(h_i)_{i\in I_{n+1}} \in \mathbb{N}^{i_{max}^{n+1}-i_{min}^{n+1}+1} : \mathbb{P}\left(\forall i \in I_{n+1}, h_i^{(M),n} = h_i \middle| \mathrm{Cond}^{(M),n,N}\right) > 0\right\}.$$

Lemma 6.2.5. *For all $M \geq 2$ such that $\left(1-\lfloor gM\rfloor M^{-1}\right)^H - \lfloor gM\rfloor^{-1} > 0$,*

$$\begin{aligned}&\sum_{n=1}^{N}\mathbb{P}\left(\bigcup_{i\in I_{n+1}}\mathrm{ErrGF}_i^{(M),n}\middle|\mathrm{Cond}^{(M),n,N}\right)\\ &\leq 2\left(N(i_{max}^0-i_{min}^0+3)+N(N+1)\right)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}.\end{aligned}$$

Proof. Let $n \in [\![1, N]\!]$. Then,

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{i\in I_{n+1}}\mathrm{ErrGF}_i^{(M),n}\middle|\mathrm{Cond}^{(M),n,N}\right)\\ &= \sum_{(h_i)_{i\in I_{n+1}}\in\mathcal{V}^{(M),n}}\mathbb{P}\left(\forall i\in I_{n+1}, h_i^{(M),n}=h_i\middle|\mathrm{Cond}^{(M),n,N}\right)\\ &\qquad\times\mathbb{P}\left(\bigcup_{i\in I_{n+1}}\mathrm{ErrGF}_i^{(M),n}\middle|\mathrm{Cond}^{(M),n,N}\cap\left\{\forall i\in I_{n+1}, h_i^{(M),n}=h_i\right\}\right)\\ &\leq \sum_{(h_i)_{i\in I_{n+1}}\in\mathcal{V}^{(M),n}}\sum_{i\in I_{n+1}}\mathbb{P}\left(\forall i'\in I_{n+1}, h_{i'}^{(M),n}=h_{i'}\middle|\mathrm{Cond}^{(M),n,N}\right)\\ &\qquad\times\mathbb{P}\left(\mathrm{ErrGF}_i^{(M),n}\middle|\mathrm{Cond}^{(M),n,N}\cap\left\{\forall i'\in I_{n+1}, h_{i'}^{(M),n}=h_{i'}\right\}\right).\end{aligned}$$

Moreover, for all $i \in I_{n+1}$ and $(h_i)_{i\in I_{n+1}} \in \mathcal{V}^{(M),n}$, by (6.2.7), if we define the event $\mathrm{Cond}^{+,(M),n,N,(h_i)_{i\in I_{n+1}}}$ as

$$\mathrm{Cond}^{+,(M),n,N,(h_i)_{i\in I_{n+1}}} := \mathrm{Cond}^{(M),n,N}\cap\left\{\forall i\in I_{n+1}, h_i^{(M),n}=h_i\right\},$$

then

$$\begin{aligned}&\mathbb{P}\left(\mathrm{ErrGF}_i^{(M),n}\middle|\mathrm{Cond}^{(M),n,N}\cap\left\{\forall i'\in I_{n+1}, h_{i'}^{(M),n}=h_{i'}\right\}\right)\\ &\leq \mathbb{1}_{\{h_i\leq H\}}\\ &\qquad\times\mathbb{P}\left(\left\{\mathrm{Card}\left(E_{i,h_i}^{(M),n}\cap G_i^{(M),n}\right)=0\right\}\cup\left\{\mathrm{Card}\left(E_{i,h_i+1}^{(M),n+1}\right)\notin W_{h_i+1}^{(M),\epsilon}\right\}\middle|\mathrm{Cond}^{+,(M),n,N,(h_i)_{i\in I_{n+1}}}\right).\end{aligned}$$

Whether $\mathrm{Card}\left(E_{i,h_i}^{(M),n}\cap G_i^{(M),n}\right)=0$ or $\mathrm{Card}\left(E_{i,h_i+1}^{(M),n+1}\right)\notin W_{h_i+1}^{(M),\epsilon}$ only depends on the number of age h_i seeds in patch i at the beginning of generation n, and not on the past dynamics, the age of

the youngest type 1 seeds in patch i as well as other patches, or the composition of other patches. Therefore, the event

$$\left\{\mathrm{Card}\left(E_{i,h_i}^{(M),n}\cap G_i^{(M),n}\right)=0\right\}\cup\left\{\mathrm{Card}\left(E_{i,h_i+1}^{(M),n+1}\right)\notin W_{h_i+1}^{(M),\epsilon}\right\}$$

is independent from most of the events whose union form $\mathrm{Cond}^{+,(M),n,N,(h_i)_{i\in I_{n+1}}}$, and

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{ErrGF}_i^{(M),n}\middle|\,\mathrm{Cond}^{(M),n,N}\cap\left\{\forall i'\in I_{n+1},h_{i'}^{(M),n}=h_{i'}\right\}\right)\\
&\leq \mathbb{1}_{\{h_i\leq H\}}\\
&\qquad\times\mathbb{P}\left(\left\{\mathrm{Card}\left(E_{i,h_i}^{(M),n}\cap G_i^{(M),n}\right)=0\right\}\cup\left\{\mathrm{Card}\left(E_{i,h_i+1}^{(M),n+1}\right)\notin W_{h_i+1}^{(M),\epsilon}\right\}\middle|\,\mathrm{Card}\left(E_{i,h_i}^{(M),n}\right)\in W_{h_i}^{(M),\epsilon}\right)\\
&\leq 2e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}
\end{aligned}$$

by Lemma 6.2.4. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i\in I_{n+1}}\mathrm{ErrGF}_i^{(M),n}\middle|\,\mathrm{Cond}^{(M),n,N}\right)\\
&\leq 2\left(i_{max}^{n+1}-i_{min}^{n+1}+1\right)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}\\
&\leq 2(i_{max}^0-i_{min}^0+1+2n+2)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{n=1}^N\mathbb{P}\left(\bigcup_{i\in I_{n+1}}\mathrm{ErrGF}_i^{(M),n}\middle|\,\mathrm{Cond}^{(M),n,N}\right)\\
&\leq\sum_{n=1}^N\left(2(2n+i_{max}^0-i_{min}^0+3)\right)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}\\
&\leq\left(2N(N+1)+2(i_0^{max}-i_0^{min})N+6N\right)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}},
\end{aligned}$$

which allows us to conclude. □

Obtaining a similar result for the quantity in (6.2.5) does not require conditioning, since $h^{(M),0}$ is determined by the initial condition.

Lemma 6.2.6. *For all $M\geq 2$ such that $\left(1-\lfloor gM\rfloor M^{-1}\right)^H-\lfloor gM\rfloor^{-1}>0$,*

$$\mathbb{P}\left(\bigcup_{i\in I_1}\mathrm{ErrGF}_i^{(M),0}\middle|\left(\overline{\mathrm{Par}}_i^{(M),n'}\right)_{\substack{0\leq n'\leq N-1\\ i\in I_{n'+1}}}\right)\leq 2\left(i_{max}^0-i_{min}^0+3\right)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}.$$

Proof. By 6.2.7 and by Lemma 6.2.4, given the initial condition,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i\in I_1}\mathrm{ErrGF}_i^{(M),0}\middle|\left(\overline{\mathrm{Par}}_i^{(M),n'}\right)_{\substack{0\le n'\le N-1\\ i\in I_{n'+1}}}\right)\\
&\le\sum_{i\in I_1}\mathbb{1}_{\{h_i^{(M),0}\le H\}}\\
&\qquad\times\mathbb{P}\left(\left\{\mathrm{Card}\left(E^{(M),0}_{i,h_i^{(M),0}}\cap G_i^{(M),0}\right)=0\right\}\cup\left\{\mathrm{Card}\left(E^{(M),1}_{i,h_i^{(M),0}+1}\right)\notin W^{(M),\epsilon}_{h_i^{(M),0}+1}\right\}\middle|\left(\overline{\mathrm{Par}}_i^{(M),n}\right)_{\substack{n'\le N-1\\ i\in I_{n'+1}}}\right)\\
&\le\sum_{i\in I_1}\mathbb{P}\left(\left\{\mathrm{Card}\left(E^{(M),0}_{i,h_i^{(M),0}}\cap G_i^{(M),0}\right)=0\right\}\cup\left\{\mathrm{Card}\left(E^{(M),1}_{i,h_i^{(M),0}+1}\right)\notin W^{(M),\epsilon}_{h_i^{(M),0}+1}\right\}\right)\\
&\le 2\left(i^0_{max}-i^0_{min}+3\right)e^{-2\epsilon^2\lfloor gM\rfloor^3((1-\lfloor gM\rfloor M^{-1})^H-\epsilon a_H)^2M^{-2}}
\end{aligned}$$

given the initial condition. □

We can now show Theorem 6.1.9 for our specific initial condition.

Proof. (Theorem 6.1.9, post-transition period) Let $M\ge 2$ be such that $\left(1-\lfloor gM\rfloor M^{-1}\right)^H-\lfloor gM\rfloor^{-1}>0$. The result is clear for $N=0$. For $N\in\mathbb{N}\backslash\{0\}$, by definition of the event $\mathrm{DiffBOA}_i^{(M),n}$,

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{n=0}^{N}\left(\left\{\forall i\in\mathbb{Z},O_i^{(M),n}=O_i^{(M),\infty,n}\right\}\cap\left\{\forall i\in\mathbb{Z},h_i^{(M),n}=h_i^{(M),\infty,n}\right\}\right)\right)\\
&=\mathbb{P}\left(\bigcap_{n=0}^{N}\bigcap_{i\in\mathbb{Z}}\overline{\mathrm{DiffBOA}}_i^{(M),n}\right)\\
&=1-\mathbb{P}\left(\bigcup_{n=0}^{N}\bigcup_{i\in\mathbb{Z}}\mathrm{DiffBOA}_i^{(M),n}\right)\\
&=1-\mathbb{P}\left(\bigcup_{n=1}^{N}\bigcup_{i\in I_n}\mathrm{DiffBOA}_i^{(M),n}\right)
\end{aligned}$$

by Eq. (6.2.1). Therefore, by Eq. (6.2.2),

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{n=0}^{N}\left(\left\{\forall i\in\mathbb{Z},O_i^{(M),n}=O_i^{(M),\infty,n}\right\}\cap\left\{\forall i\in\mathbb{Z},h_i^{(M),n}=h_i^{(M),\infty,n}\right\}\right)\right)\\
&\ge 1-\mathbb{P}\left(\bigcup_{n=0}^{N-1}\bigcup_{i\in I_{n+1}}\mathrm{Par}_i^{(M),n}\cup\mathrm{ErrG}_i^{(M),n}\cup\mathrm{ErrF}_i^{(M),n+1}\right),
\end{aligned}$$

which can be bounded from below by $1-$ Eq.(6.2.4) $-$ Eq.(6.2.5) $-$ Eq.(6.2.6). By Lemmas 6.2.2, 6.2.6 and 6.2.5, we can show that each of the three terms converges to 0 when $M\to+\infty$, allowing us to conclude. □

We now explain how to generalize the proof of Theorem 6.1.9 to a more general sequence of initial conditions not necessarily satisfying (IC1) and (IC2). Since $(\xi^{(M),n},h^{(M),n})_{n\in\mathbb{N}}$ is Markovian for all $M\ge 2$, it is sufficient to show that the process does not deviate from the BOA process during the transition period, and that at the end of this period, $(\xi^{(M),H+1},h^{(M),H+1})$ satisfies (IC1) and (IC2).

In order to do so, let $n\in[\![0,H]\!]$ be a generation from the transition period, and let $i\in\mathbb{Z}$. We distinguish three cases.

1. If $O_i^\infty = 0$, then by condition (A) patch i is initially empty, so all viable $h_i^{\infty,0}$ seeds it contains are of type 0. We are then in the same situation as during the post-transition period.

2. If $O_i^\infty = 1$ and $h_i^{(M),n} > n-1$, then the age $h_i^{(M),n}$ seeds in patch i were already present initially. Similarly as before and using condition (C), we can show that with high probability, the number of remaining type 1 seeds of age $h_i^{(M),n}$ is roughly equal to $g_i(1-g)^n M$, and at least one of them germinates during generation n.

3. If $O_i^\infty = 1$ and $h_i^{(M),n} \leq n+1$, then the age $h_i^{(M),n}$ seeds in patch i were not present initially. By Lemma 6.2.2, they are all of the same type with high probability, and we are back to the case considered during the post-transition period.

6.3 Extinction threshold for the k-parent WFSB metapopulation process

This section is devoted to the proof of Theorem 6.1.10, that is, to the proof of the existence of a critical extinction probability $p_{crit}(H)$ depending only on the maximal dormancy duration H. In order to do so, we will first formalize the coupling between the k-parent WFSB metapopulation process and a BOA process. Then, we will explain how the issue of occupied patches in the BOA process can be seen as a percolation problem. We will conclude using a specific case of Eq.(4) in [HS21].

6.3.1 Coupling between the k-parent WFSB metapopulation process and the BOA process

In all that follows, let $(\xi, h) \in \mathcal{F}_M \times \mathcal{H}_M$, let $(\xi^n, h^n)_{n\in\mathbb{N}}$ be the k-parent WFSB metapopulation process with parameters (M, H, g, c, p) and initial condition (ξ, h), and let $(O^{k,n}, h^{k,n})_{n\in\mathbb{N}}$ be the associated k-parent occupancy process. In order to couple a BOA process to $(\xi^n, h^n)_{n\in\mathbb{N}}$, for all $n \in \mathbb{N}^*$, we denote by $(\mathrm{Ext}_i^n)_{i\in\mathbb{Z}}$ the extinction events used to define (ξ^n, h^n) given (ξ^{n-1}, h^{n-1}). In other words, for all $n \in \mathbb{N}^*$ and $i \in \mathbb{Z}$, $\mathrm{Ext}_i^n = 1$ if, and only if the patch i was extinct during the n-th generation. We then define the coupled BOA process $(O^{\infty,n}, h^{\infty,n})_{n\in\mathbb{N}}$ as the BOA process with parameters (H, p) and initial condition $(O^{k,0}, h^{k,0})$, constructed using the extinction events $(\mathrm{Ext}_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}^*}$: for all $n \in \mathbb{N}$, $(O^{\infty,n+1}, h^{\infty,n+1})$ is constructed using $(O^{\infty,n}, h^{\infty,n})$ and the extinction events $(\mathrm{Ext}_i^{n+1})_{i\in\mathbb{Z}}$. This coupling satisfies the following property, whose proof is postponed until later in this section for the sake of clarity.

Proposition 6.3.1. *For all $n \in \mathbb{N}$ and $i \in \mathbb{Z}$,*

$$O_i^{k,n} \leq O_i^{\infty,n} \text{ and } h_i^{k,n} \geq h_i^{\infty,n}.$$

Therefore, at any generation $n \in \mathbb{N}$, the set of patches which contain nonexpired seeds in the k-parent WFSB metapopulation process is included in the set of reachable patches in the BOA process (that is, patches i such that $O_i^{\infty,n}\mathbb{1}_{\{h_i^{\infty,n}\leq H\}} = 1$). In particular, a consequence of this coupling is the following corollary.

Corollary 6.3.2. *For all $n \in \mathbb{N}$,*

$$\mathbb{P}\left(1 - \prod_{(i,j)\in\mathbb{Z}\times[\![1,M]\!]} \left(1 - \mathbb{1}_{\{h_{i,j}^n \leq H\}}\xi_{i,j}^n\right) = 1\right) \leq \mathbb{P}\left(1 - \prod_{i\in\mathbb{Z}} \left(1 - \mathbb{1}_{\{h_i^{\infty,n}\leq H\}}O_i^{\infty,n}\right) = 1\right).$$

Proof. Let $n \in \mathbb{N}$. By definition of the k-parent occupancy process, for all $(i,j) \in \mathbb{Z} \times [\![1, M]\!]$,

$$\xi_{i,j}^n \leq O_i^{k,n}.$$

Indeed, both $\xi_{i,j}^n$ and $O_i^{k,n}$ are $\{0,1\}$-valued, and if $\xi_{i,j}^n = 1$, then $O_i^{k,n} = 1$.

Moreover, if $O_i^{k,n} = 1$, then $h_i^{k,n}$ is the age of the youngest type 1 seed in patch i. Therefore, for all $(i,j) \in \mathbb{Z} \times [\![1, M]\!]$, if $\xi_{i,j}^n = 1$, then $h_{i,j}^n \geq h_i^{k,n}$. We deduce that

$$\mathbb{1}_{\{h_{i,j}^n \leq H\}} \xi_{i,j}^n \leq \mathbb{1}_{\{h_{i,j}^{k,n} \leq H\}} O_i^{k,n}.$$

By Proposition 6.3.1, we obtain

$$\mathbb{1}_{\{h_{i,j}^n \leq H\}} \xi_{i,j}^n \leq \mathbb{1}_{\{h_i^{\infty,n} \leq H\}} O_i^{\infty,n}.$$

Taking the product over all $(i,j) \in \mathbb{Z} \times [\![1, M]\!]$ yields

$$\begin{aligned} 1 - \prod_{(i,j) \in \mathbb{Z} \times [\![1,M]\!]} \left(1 - \mathbb{1}_{\{h_{i,j}^n \leq H\}} \xi_{i,j}^n\right) \leq & 1 - \prod_{(i,j) \in \mathbb{Z} \times [\![1,M]\!]} \left(1 - \mathbb{1}_{\{h_i^{\infty,n} \leq H\}} O_i^{\infty,n}\right) \\ \leq & 1 - \prod_{i \in \mathbb{Z}} \left(1 - \mathbb{1}_{\{h_i^{\infty,n} \leq H\}} O_i^{\infty,n}\right). \end{aligned}$$

since all the terms of the product are $\{0,1\}$-valued, and we can conclude. □

We now show Proposition 6.3.1.

Proof. (Proposition 6.3.1) We show the result by induction. For $n = 0$, since $(O^{\infty,0}, h^{\infty,0}) = (O^{k,0}, h^{k,0})$,

$$\begin{aligned} 1 - \prod_{(i,j) \in \mathbb{Z} \times [\![1,M]\!]} \left(1 - \mathbb{1}_{\{h_{i,j}^0 \leq H\}} \xi_{i,j}^0\right) &= 1 - \prod_{i \in \mathbb{Z}} \prod_{j \in [\![1,M]\!]} \left(1 - \mathbb{1}_{\{h_{i,j}^0 \leq H\}} \xi_{i,j}^0\right) \\ &= 1 - \prod_{i \in \mathbb{Z}} \left(1 - \mathbb{1}_{\{h_i^{k,0} \leq H\}} O_i^{k,0}\right) \\ &= 1 - \prod_{i \in \mathbb{Z}} \left(1 - \mathbb{1}_{\{h_i^{\infty,0} \leq H\}} O_i^{\infty,n}\right), \end{aligned}$$

so the result is true for $n = 0$.

Let then $n \in \mathbb{N}$, and we assume that for all $i \in \mathbb{Z}$,

$$O_i^{k,n} \leq O_i^{\infty,n} \text{ and } h_i^{k,n} \geq h_i^{\infty,n}.$$

Let $i \in \mathbb{Z}$. We first show that $O_i^{k,n+1} \leq O_i^{\infty,n+1}$. Since $O_i^{k,n+1} \in \{0,1\}$, if $O_i^{\infty,n+1} = 1$, then $O_i^{k,n+1} \leq O_i^{\infty,n+1}$. Therefore, we assume $O_i^{\infty,n+1} = 0$. Notice that by definition of the BOA process, $(O_i^{\infty,n})_{n \in \mathbb{N}}$ is an increasing sequence. Indeed, for all $n \geq 0$ and $i \in \mathbb{Z}$, $O_i^{\infty,n+1}$ is set equal to $O_i^{\infty,n}$ or 1, so if $O_i^{\infty,n} = 1$, then $O_i^{\infty,n+1} \in \{O_i^{\infty,n}, 1\} = \{1\}$, and for all $n' \geq n$, we have $O_i^{\infty,n'} = 1$. This means that $O_i^{\infty,n+1} = 0$ implies $O_i^{\infty,n} = 0$ and $O_i^{k,n} = 0$. Moreover, it also means that both neighbouring patches were either extinct or not reachable in generation n. We deduce

$$\begin{aligned} \left(1 - \mathrm{Ext}_{i+1}^{n+1}\right) O_{i+1}^{\infty,n} \mathbb{1}_{\{h_{i+1}^{\infty,n} \leq H\}} &= 0 \\ \text{and } \left(1 - \mathrm{Ext}_{i-1}^{n+1}\right) O_{i-1}^{\infty,n} \mathbb{1}_{\{h_{i-1}^{\infty,n} \leq H\}} &= 0. \end{aligned}$$

Therefore, by the induction hypothesis,

$$\begin{aligned}\left(1-\mathrm{Ext}_{i+1}^{n+1}\right) O_{i+1}^{k,n}\mathbb{1}_{\{h_{i+1}^{k,n}\leq H\}}&=0\\ \text{and } \left(1-\mathrm{Ext}_{i-1}^{n+1}\right) O_{i-1}^{k,n}\mathbb{1}_{\{h_{i-1}^{k,n}\leq H\}}&=0,\end{aligned}$$

which means that the patches $i-1$ and $i+1$ are either extinct or containing only ghost type 0 seeds. Combined with the knowledge that $O_i^{k,n}=0$, we obtain that $O_i^{k,n+1}=0$.

We now have to show that $h_i^{k,n+1}\geq h_i^{\infty,n+1}$. Since $h_i^{k,n}\geq h_i^{\infty,n}$ and since $h_i^{k,n+1}$ (resp. $h_i^{\infty,n+1}$) is either equal to $h_i^{k,n}+1$ (resp. $h_i^{\infty,n}+1$) or equal to 0, the only potential issue is when $h_i^{k,n+1}=0$. Let us assume that $h_i^{k,n+1}=0$. This means that new seeds were just produced, and implies that

$$1-\prod_{i'=i-1}^{i+1}\left(1-\left(1-\mathrm{Ext}_{i'}^{n+1}\right) O_{i'}^{k,n}\mathbb{1}_{\{h_{i'}^{k,n}\leq H\}}\right)=1,$$

i.e, that non-expired seeds were present in at least one of the patches $\{i-1,\, i,\, i+1\}$, and that at least one of these patches was not affected by an extinction event. Moreover, if

$$\prod_{i'=i-1}^{i+1}\left(1-\left(1-\mathrm{Ext}_{i'}^{n+1}\right) O_{i'}^{\infty,n}\mathbb{1}_{\{h_{i'}^{\infty,n}\leq H\}}\right)=0,$$

then $h_i^{\infty,n+1}=0$. Using the induction hypothesis yields

$$\begin{aligned}\prod_{i'=i-1}^{i+1}\left(1-\left(1-\mathrm{Ext}_{i'}^{n+1}\right) O_{i'}^{\infty,n}\mathbb{1}_{\{h_{i'}^{\infty,n}\leq H\}}\right)&\leq\prod_{i'=i-1}^{i+1}\left(1-\left(1-\mathrm{Ext}_{i'}^{n+1}\right) O_{i'}^{k,n}\mathbb{1}_{\{h_{i'}^{k,n}\leq H\}}\right)\\&=0,\end{aligned}$$

hence $h_i^{\infty,n+1}=0=h_i^{k,n+1}$ and we can conclude. □

6.3.2 Percolation problem

In order to show Theorem 6.1.10, we now link the BOA process to a percolation problem. More specifically, we rephrase the question of which patches are reachable in the BOA process as an oriented site percolation problem. Indeed, we can see patch $i\in\mathbb{Z}$ in generation $n\in\mathbb{N}$ as the site (i,n) of the space $\mathbb{Z}\times\mathbb{N}$. Each site $(i,n)\in\mathbb{Z}\times\mathbb{N}$ is *open* (the analog of *non-extinct* in the terminology of percolation) with probability $1-p$, and *closed* (i.e, extinct) otherwise. Reachable patches can be seen as sites of the space $\mathbb{Z}\times\mathbb{N}$ linked to a site of $\mathbb{Z}\times\{0\}$ by a path of open sites

$$(i_0,n_0)=(i_0,0)\longrightarrow(i_1,n_1)\longrightarrow\ldots\longrightarrow(i_L,n_L)=(i,n)$$

such that $O_{i_0}^{\infty,0}\times\mathbb{1}_{\{h_{i_0}^{\infty,n}\leq H\}}=1$, $i_1\in\{i_0-1,i_0,i_0+1\}$, $n_1-n_0\in[\![1,H-h_{i_0}^{\infty,n}+1]\!]$, and for all $l\in[\![2,L]\!]$,

$$i_l\in\{i_{l-1}-1,i_{l-1},i_{l-1}+1\}\qquad\text{and}\qquad n_l-n_{l-1}\in[\![1,H+1]\!].\tag{6.3.1}$$

For all $n\in\mathbb{N}$, let $S_n(p)$ be the set of all the sites (i,n) with $i\in\mathbb{Z}$ that are connected to $(0,0)$ by a path of open sites satisfying $i_1\in\{i_0-1,i_0,i_0+1\}$, $n_1-n_0\in[\![1,H+1]\!]$ and (6.3.1). Equivalently, let $(O^{\{0\},n},h^{\{0\},n})_{n\in\mathbb{N}}$ be the BOA process with parameters (H,p) and initial condition satisfying:

1. $O_0^{\{0\},0}=1$ and $h_0^{\{0\},0}=0$.

2. For all $i \in \mathbb{Z}\backslash\{0\}$, $O_i^{\{0\},0} = 0$ and $h_i^{\{0\},0} = 0$.

We can then define $S_n(p)$ as

$$S_n(p) := \left\{ i \in \mathbb{Z} : O_i^{\{0\},n} \mathbb{1}_{\{h_i^{\{0\},n} \leq H\}} = 1 \right\}.$$

Under this notation, a direct consequence of Eq. (4) in [HS21] is the following proposition.

Proposition 6.3.3. *There exists a unique* $p_{crit}(H) \in (0,1)$ *such that*

$$\begin{aligned} &\forall p \in [0, p_{crit}(H)), \mathbb{P}\left(\forall n \in \mathbb{N}, S_n(p) \neq \emptyset\right) > 0 \\ \textit{and} \quad &\forall p \in (p_{crit}(H), 1], \mathbb{P}\left(\forall n \in \mathbb{N}, S_n(p) \neq \emptyset\right) = 0. \end{aligned}$$

What remains to show is that $p_{crit}(H)$ is indeed the extinction threshold we are looking for.

6.3.3 Proof of Theorem 6.1.10

In order to prove Theorem 6.1.10, we make three observations. First, for all $n \in \mathbb{N}$, the event $\{S_n(p) \neq \emptyset\}$ is the same as the event

$$\left\{ 1 - \prod_{i \in \mathbb{Z}} \left(1 - O_i^{\{0\},n} \mathbb{1}_{\{h_i^{\{0\},n} \leq H\}}\right) = 1 \right\}.$$

Moreover, for all finite subsets $\mathcal{L}$ of $\mathbb{Z}$, let $(O^{\mathcal{L}}, h^{\mathcal{L}}) \in \mathcal{F}^\infty \times \mathcal{H}^\infty$ satisfy the two following conditions:

- For all $i \in \mathcal{L}$, $O_i^{\mathcal{L}} = 1$ and $h_i^{\mathcal{L}} = 0$.
- For all $i \in \mathbb{Z}\backslash\mathcal{L}$, $O_i^{\mathcal{L}} = 0$ and $h_i^{\mathcal{L}} = 0$.

Let also $(O^{\mathcal{L},n}, h^{\mathcal{L},n})_{n\in\mathbb{N}}$ be the BOA process with parameters (H,p) and initial condition $(O^{\mathcal{L}}, h^{\mathcal{L}})$. That is, $(O^{\mathcal{L},n}, h^{\mathcal{L},n})_{n\in\mathbb{N}}$ is the BOA process starting from the state where all the patches in $\mathcal{L}$ are of type 1 and all the patches in $\mathcal{L}^c$ of type 0. Notice that if $\mathcal{L} = \{0\}$, then the definition of $\left(O^{\{0\},n}, h^{\{0\},n}\right))_{n\in\mathbb{N}}$ matches the one given above. We then have the following result.

Lemma 6.3.4. *For all finite subset* $\mathcal{L}$ *of* $\mathbb{Z}$ *and for all* $n \in \mathbb{N}$*,*

$$\mathbb{P}\left(1 - \prod_{i \in \mathbb{Z}} \left(1 - O_i^{\mathcal{L},n} \mathbb{1}_{\{h_i^{\mathcal{L},n} \leq H\}}\right) = 0\right) \geq \mathbb{P}\left(1 - \prod_{i \in \mathbb{Z}} \left(1 - O_i^{\{0\},n} \mathbb{1}_{\{h_i^{\{0\},n} \leq H\}}\right) = 0\right)^{\mathrm{Card}(\mathcal{L})}.$$

This lemma gives a lower bound of the probability that no patches are reachable in at least n generations in the BOA process starting from the patches in $\mathcal{L}$, each one of them containing type 1 seeds of age 0. This lower bound involves the probability that no patches are reachable in at least n generations starting from *only one patch*, which is used in the definition of $p_{crit}(H)$.

Proof. Let $n \in \mathbb{N}$ and let $\mathcal{L}$ be a finite subset of $\mathbb{Z}$. First, we observe that if we couple all the BOA processes considered by constructing them using the same extinction events,

$$1 - \prod_{i \in \mathbb{Z}} \left(1 - O_i^{\mathcal{L},n} \mathbb{1}_{\{h_i^{\mathcal{L},n} \leq H\}}\right) = 1 - \prod_{i' \in \mathcal{L}} \left[\prod_{i \in \mathbb{Z}} \left(1 - O_i^{\{i'\},n} \mathbb{1}_{\{h_i^{\{i'\},n} \leq H\}}\right)\right]. \tag{6.3.2}$$

Indeed, each one of the reachable patches in the BOA process with initial conditions $(O^{\mathcal{L}}, h^{\mathcal{L}})$ is connected by a path of nonextinct patches to a patch in $\mathcal{L}$, and so there exists $i_0 \in \mathcal{L}$ such as the

patch is also reachable in the BOA process with initial condition $(O^{\{i_0\}}, h^{\{i_0\}})$. We can then use the fact that all the quantities appearing in the product are $\{0,1\}$-valued.

Moreover, for $i_0, i_1 \in \mathcal{L}$ and again using our coupling, knowing that no patch is reachable in n generations starting from i_0 increases the probability that no patch is reachable in n generations starting from i_1. Indeed, informally, the fact that no patch is reachable starting from i_0 "blocks" some patches, which cannot be used by a path linking i_1 to other patches. In a more formal way, we can rewrite the event

$$\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_1\},n} \mathbb{1}_{\{h_i^{\{i_1\},n} \leq H\}}\right) = 0$$

as

$$\exists i^{(0)}, ..., i^n \in \mathbb{Z} \text{ satisfying Condition (C)},$$

where Condition (C) is defined as

- $i^{(0)} = i_1$,
- $\forall n' \in [\![0, n-1]\!], |i^{(n'+1)} - i^{(n')}| \leq 1$,
- $\forall n' \in [\![0, n]\!], O_{i^{(n')}}^{\{i_1\},n'} \mathbb{1}_{\{h_{i^{(n')}}^{\{i_1\},n'} \leq H\}} = 1$,
- $\forall n' \in [\![0, n-1]\!], i^{(n'+1)} \neq i^{(n')} \Longrightarrow h_{i^{(n'+1)}}^{\{i_1\},n'+1} = 0$.

Using this observation,

$$\begin{aligned}
&\mathbb{P}\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_1\},n} \mathbb{1}_{\{h_i^{\{i_1\},n} \leq H\}}\right) = 0 \middle| \prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right) \\
&= \mathbb{P}\left(\exists i^{(0)}, ..., i^{(n)} \text{ satisfying (C)} \middle| \prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right) \\
&= \frac{\mathbb{P}\left(\left(\exists i^{(0)}, ..., i^{(n)} \text{ satisfying (C)}\right) \bigcup \left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right)\right)}{\mathbb{P}\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right)} \\
&= \frac{\mathbb{P}\left(\exists i^{(0)}, ..., i^{(n)} \text{ satisfying (C)}\right) \mathbb{P}\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1 \middle| \exists i^{(0)}, ..., i^{(n)} \text{ satisfying (C)}\right)}{\mathbb{P}\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right)}
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathbb{P}\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1 \middle| \exists i^{(0)}, ..., i^{(n)} \text{ satisfying (C)}\right) \\
&= \mathbb{P}\left(\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right)\right. \\
&\qquad \left.\bigcap \left(\forall n' \in [\![0, n-1]\!], \text{ if } h_{i^{(n'+1)}}^{\{i_1\},n'+1} = 0, \text{ then } \forall \tilde{i} \in [\![-1, 1]\!], O_{i^{(n')}+\tilde{i}}^{\{i_0\},n'} \mathbb{1}_{\{h_{i^{(n')}+\tilde{i}}^{\{i_0\},n'} \leq H\}} = 0\right)\right) \\
&\leq \mathbb{P}\left(\prod_{i\in\mathbb{Z}} \left(1 - O_i^{\{i_0\},n} \mathbb{1}_{\{h_i^{\{i_0\},n} \leq H\}}\right) = 1\right).
\end{aligned}$$

which means that

$$\mathbb{P}\left(\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_1\},n}\mathbb{1}_{\{h_i^{\{i_1\},n}\leq H\}}\right)=0\middle|\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_0\},n}\mathbb{1}_{\{h_i^{\{i_0\},n}\leq H\}}\right)=1\right)$$
$$\leq\mathbb{P}\left(\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_1\},n}\mathbb{1}_{\{h_i^{\{i_1\},n}\leq H\}}\right)=0\right),$$

or, if we consider the complementary events,

$$\mathbb{P}\left(\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_1\},n}\mathbb{1}_{\{h_i^{\{i_1\},n}\leq H\}}\right)=1\middle|\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_0\},n}\mathbb{1}_{\{h_i^{\{i_0\},n}\leq H\}}\right)=1\right)$$
$$\geq\mathbb{P}\left(\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_1\},n}\mathbb{1}_{\{h_i^{\{i_1\},n}\leq H\}}\right)=1\right).$$

Hence, for $i_0\in\mathcal{L}$, by Eq. (6.3.2),

$$\begin{aligned}\mathbb{P}\left(1-\prod_{i\in\mathbb{Z}}\left(1-O_i^{\mathcal{L},n}\mathbb{1}_{\{h_i^{\mathcal{L},n}\leq H\}}\right)=0\right)&=\mathbb{P}\left(1-\prod_{i'\in\mathcal{L}}\left[\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i'\},n}\mathbb{1}_{\{h_i^{\{i'\},n}\leq H\}}\right]\right)=0\right)\\&=\mathbb{P}\left(\bigcap_{i'\in\mathcal{L}}\left\{\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i'\},n}\mathbb{1}_{\{h_i^{\{i'\},n}\leq H\}}\right)=1\right\}\right)\\&\geq\mathbb{P}\left(\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{i_0\},n}\mathbb{1}_{\{h_i^{\{i_0\},n}\leq H\}}\right)=1\right)^{\mathrm{Card}(\mathcal{L})}\\&=\mathbb{P}\left(\prod_{i\in\mathbb{Z}}\left(1-O_i^{\{0\},n}\mathbb{1}_{\{h_i^{\{0\},n}\leq H\}}\right)=1\right)^{\mathrm{Card}(\mathcal{L})},\end{aligned}$$

where the invariance by translation of the process is used to pass from the last but first to the last line. □

We recall that the k-parent occupancy process associated to $(\xi^n,h^n)_{n\in\mathbb{N}}$ is denoted by

$$(O^{k,n},h^{k,n})_{n\in\mathbb{N}}.$$

The coupling based on the extinction events also yields the following lemma.

Lemma 6.3.5. *Let $\mathcal{L}\subset\mathbb{Z}$ be the set defined as*

$$\mathcal{L}:=\left\{i\in\mathbb{Z}:O_i^{k,0}=1\right\}.$$

Then,

$$\mathbb{P}\left(1-\prod_{i\in\mathbb{Z}}\left(1-O_i^{k,n}\mathbb{1}_{\{h_i^{k,n}\leq H\}}\right)=0\right)\geq\mathbb{P}\left(1-\prod_{i\in\mathbb{Z}}\left(1-O_i^{\mathcal{L},n}\mathbb{1}_{\{h_i^{\mathcal{L},n}\leq H\}}\right)=0\right).$$

Indeed, if $(\xi^n,h^n)_{n\in\mathbb{N}}$ (hence $(O^{k,n},h^{k,n})_{n\in\mathbb{N}}$) and $(O^{\mathcal{L},n},h^{\mathcal{L},n})_{n\in\mathbb{N}}$ are constructed using the same extinction events, then all the patches occupied by the k-parent WFSB metapopulation process are also reachable by the BOA process $(O^{\mathcal{L},n},h^{\mathcal{L},n})_{n\in\mathbb{N}}$. Here deviations from the BOA process $(O^{\mathcal{L},n},h^{\mathcal{L},n})_{n\in\mathbb{N}}$ can also occur if the youngest type 1 seeds in (ξ^0,h^0) are *not* of age 0, but older.

We can now prove Theorem 6.1.10.

Proof. (Theorem 6.1.10) Let $p_{crit}(H)$ be given by Proposition 6.3.3. We assume that $p > p_{crit}(H)$. Let also $n \in \mathbb{N}$, and let $\mathcal{L} \subset \mathbb{Z}$ be defined as in Lemma 6.3.5.

By Lemma 6.3.5,

$$\begin{aligned}\mathbb{P}\left(\forall i \in \mathbb{Z}, O_i^{k,n}\mathbb{1}_{\{h_i^{k,n}\leq H\}} = 0\right) &= \mathbb{P}\left(1 - \prod_{i\in\mathbb{Z}}\left(1 - O_i^{k,n}\mathbb{1}_{\{h_i^{k,n}\leq H\}}\right) = 0\right) \\ &\geq \mathbb{P}\left(1 - \prod_{i\in\mathbb{Z}}\left(1 - O_i^{\mathcal{L},n}\mathbb{1}_{\{h_i^{\mathcal{L},n}\leq H\}}\right) = 0\right).\end{aligned}$$

Using Lemma 6.3.4, we obtain

$$\begin{aligned}\mathbb{P}\left(\forall i \in \mathbb{Z}, O_i^{k,n}\mathbb{1}_{\{h_i^{k,n}\leq H\}} = 0\right) &\geq \mathbb{P}\left(1 - \prod_{i\in\mathbb{Z}}\left(1 - O_i^{\{0\},n}\mathbb{1}_{\{h_i^{\{0\},n}\leq H\}}\right) = 0\right)^{\mathrm{Card}(\mathcal{L})} \\ &= \mathbb{P}\left(S_n(p) = \emptyset\right)^{\mathrm{Card}(\mathcal{L})}.\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{n\to+\infty}\mathbb{P}\left(\forall i \in \mathbb{Z}, O_i^{k,n}\mathbb{1}_{\{h_i^{k,n}\leq H\}} = 0\right) &\geq \lim_{n\to+\infty}\mathbb{P}(S_n(p) = \emptyset)^{\mathrm{Card}(\mathcal{L})} \\ &\geq 1\end{aligned}$$

by Proposition 6.3.3, and we can conclude. □

6.4 Appendix - Computation of $p_{crit}(H)$

In this section, we briefly explain how to compute $p_{crit}(H)$, and how to implement this approach and obtain an approximation for $p_{crit}(H)$. The computation method is a direct adaptation of Section 3 in [Dur84]. Our goal here is not to obtain very precise approximations, but rather to have a rough estimate of $p_{crit}(H)$, and use it to assess the impact of the presence of a seed bank on the extinction threshold.

We first introduce the following notation. For all $i \in \mathbb{Z}$ and $n \in \mathbb{N}$, let $U^{i,n}$ be a random variable such that $U^{i,n} \sim \mathrm{Unif}([0,1])$. We assume that all the random variables $(U^{i,n})_{i\in\mathbb{Z},n\in\mathbb{N}}$ are independent. For all $p \in [0,1]$, let $\mathcal{S}_p$ be the set defined as

$$\mathcal{S}_p := \left\{(i,n) : i \in \mathbb{Z}, n \in \mathbb{N} \text{ and } U^{i,n} \geq p\right\}.$$

$\mathcal{S}_p$ can be interpreted as the set of patches which would be non-extinct, if the extinction probability was equal to p.

For all $x, y \in \mathbb{Z}$, $n^{(x)}, n^{(y)} \in \mathbb{N}$, $H \in \mathbb{N}$ and $p \in [0,1]$, we will say that $(x, n^{(x)})$ is *(H,p)-reachable from* $(y, n^{(y)})$, and denote it as $(y, n^{(y)}) \xrightarrow[(H,p)]{} (x, n^{(x)})$, if there exists $L \in \mathbb{N}$, x_0, x_1,..., $x_L \in \mathbb{Z}$ and n_0, n_1,..., $n_L \in \mathbb{N}$ such that:

1. $x_0 = y$, $n_0 = n^{(y)}$, $x_l = x$ and $n_l = n^{(x)}$,
2. $\forall l \in [\![1, L]\!]$, $x_l \in \{x_{l-1} - 1, x_{l-1}, x_{l-1} + 1\}$ and $1 \leq n_l - n_{l-1} \leq H + 1$,
3. $\forall l \in [\![1, L]\!]$, $(x_l, n_l) \in \mathcal{S}_p$.

In other words, $(y, n^{(y)}) \xrightarrow[(H,p)]{} (x, n^{(x)})$ if there exists a path of open sites going from $(y, n^{(y)})$ to $(x, n^{(x)})$, spending at most H generations in each patch.

Moreover, for all $p \in [0,1]$ and $n \in \mathbb{N}$, let $\bar{\xi}_{\mathbf{n}}(H,p)$ be the set defined as:

$$\bar{\xi}_{\mathbf{n}}(H,p) := \left\{ x \in \mathbb{Z} : \exists h_x, h_y \in [\![0,H]\!], \exists y \in \mathbb{Z}\backslash(\mathbb{N}\backslash\{0\}), (y,h_y) \xrightarrow[(H,p)]{} (x, n+h_x) \right\},$$

and let $\bar{r}_n(H,p) := \sup \bar{\xi}_{\mathbf{n}}(H,p)$. $\bar{\xi}_{\mathbf{n}}(H,p)$ is akin to the set of patches which are reachable in n generations in a BOA process with parameters (H,p), but starting from an infinite number of patches.

A direct adaptation of Section 3 from [Dur84] yields the following result.

Lemma 6.4.1. *For all $H \in \mathbb{N}$,*

$$p_{crit}(H) := \max \left\{ p \in [0,1] : \lim_{n \to +\infty} \frac{\bar{r}_n(H,p)}{n} \geq 0 \right\}.$$

Therefore, in order to compute $p_{crit}(H)$, it is possible to simulate the random variable $\bar{r}_n(H,p)$ for a large value of n and for different values of p.

Let $H \in \mathbb{N}$. In order to obtain an approximation for $p_{crit}(H)$, we first define some approximations for $\bar{\xi}_{\mathbf{n}}(H,p)$ and $\bar{r}_n(H,p)$. Let $p \in [0,1]$. For all $x, y \in [\![-10500, 10500]\!]$ and $n^{(x)}, n^{(y)} \in [\![0, 10000]\!]$, we will say that $(x, n^{(x)})$ is *approximatively (H,p)-reachable from* $(y, n^{(y)})$, and denote it as

$$(y, n^{(y)}) \xrightarrow[\mathrm{Approx}(H,p)]{} (x, n^{(x)}),$$

if there exists $L \in \mathbb{N}$, x_0,... , $x_L \in [\![-10500, 10500]\!]$ and n_0,... ,$n_L \in [\![0, 10000]\!]$ such that:

1. $x_0 = y$, $n_0 = n^{(y)}$, $x_l = x$ and $n_l = n^{(x)}$.
2. $\forall l \in [\![1, L]\!]$, $x_l \in \{x_{l-1} - 1, x_{l-1}, x_{l-1} + 1\}$ and $1 \leq n_l - n_{l-1} \leq H + 1$.
3. $\forall l \in [\![1, L]\!]$, if $x_l \neq -10500$, then $(x_l, n_l) \in \mathcal{S}_p$ and $x_l \neq 10500$.

Therefore, in the approximation, the paths linking two sites together have to remain in the domain $[\![-10500, 10500]\!]$, with extra conditions at the border of the domain. Since the value of the quantity we are interested in depends on the presence of paths staying close to the centre of the domain, we can assume that the border conditions chosen will not affect the approximate value.

We then define

$$\mathrm{Approx}(\bar{\xi}_{\mathbf{n}}(H,p)) := \left\{ x \in \mathbb{Z} : \exists h_x, h_y \in [\![0,H]\!], \exists y \in \mathbb{Z}\backslash(\mathbb{N}\backslash\{0\}), (y,h_y) \xrightarrow[\mathrm{Approx}(H,p)]{} (x, n+h_x) \right\},$$

and let $\mathrm{Approx}(\bar{r}_n(H,p)) := \sup \mathrm{Approx}(\bar{\xi}_{\mathbf{n}}(H,p))$.

In order to compute an approximate value for $p_{crit}(H)$, we apply the following method, starting from $p = 0.99$.

1. We simulate the random variable $\mathrm{Approx}(\bar{r}_{10000}(H,p)) \times (10000)^{-1}$.
2. If the value obtained is larger than -0.005, we take $p_{crit}(H) = p$.
3. Otherwise, we substitute p with $p - 0.01$, and restart at Step 1.

Bibliography

[ABH00] P. Alpert, E. Bone, and C. Holzapfel. “Invasiveness, invasibility and the role of environmental stress in the spread of non-native plants”. In: *Perspectives in Plant Ecology, Evolution and Systematics* 3.1 (2000), pp. 52–66.

[AD15] S.E. Alm and M. Deijfen. “First Passage Percolation on Z2: A Simulation Study”. In: *Journal of Statistical Physics* 161.3 (2015), pp. 657–678.

[ADH17] A. Auffinger, M. Damron, and J. Hanson. *50 years of first-passage percolation*. Vol. 68. American Mathematical Society, 2017.

[AN72] K.B. Athreya and P.E. Ney. *Branching Processes*. Springer Berlin Heidelberg, 1972.

[AN94] F.R. Adler and B. Nuernberger. “Persistence in patchy irregular landscapes”. In: *Theoretical Population Biology* 45.1 (1994), pp. 41–75.

[AP01] P. Amarasekare and H. Possingham. “Patch dynamics and metapopulation theory: the case of successional species”. In: *Journal of Theoretical Biology* 209.3 (2001), pp. 333–344.

[AP02] S.E. Alm and R. Parviainen. “Lower and upper bounds for the time constant of first-passage percolation”. In: *Combinatorics, Probability and Computing* 11.5 (2002), pp. 433–445.

[APR05] M.B. Araújo, R.G. Pearson, and C. Rahbek. “Equilibrium of species’ distributions with climate”. In: *Ecography* 28.5 (2005), pp. 693–695.

[Ara+08] M.B. Araújo et al. “Quaternary climate changes explain diversity among reptiles and amphibians”. In: *Ecography* 31.1 (2008), pp. 8–15.

[Asm08] S. Asmussen. *Applied probability and queues*. Vol. 51. Springer Science & Business Media, 2008.

[Aus+97] F. Austerlitz et al. “Evolution of coalescence times, genetic diversity and structure during colonization”. In: *Theoretical Population Biology* 51.2 (1997), pp. 148–164.

[Bag04] M. Baguette. “The classical metapopulation theory and the real, natural world: a critical appraisal”. In: *Basic and Applied Ecology* 5.3 (2004), pp. 213–224.

[Bar+13] N.H. Barton et al. “Genetic hitchhiking in spatially extended populations”. In: *Theoretical Population Biology* 87 (2013), pp. 75–89.

[Bat+14] D. Bates et al. “Fitting linear mixed-effects models using lme4”. In: *arXiv preprint arXiv:1406.5823* (2014).

[BB14] C. Baskin and J. Baskin. *Seeds: Ecology, biogeography, and evolution of dormancy and germination*. Elsevier, 2014.

[BBR19] T.M. Blackburn, C. Bellard, and A. Ricciardi. “Alien versus native species as drivers of recent extinctions”. In: *Frontiers in Ecology and the Environment* 17.4 (2019), pp. 203–207.

[BCB16] C. Bellard, P. Cassey, and T.M. Blackburn. “Alien species as a driver of recent extinctions”. In: *Biology letters* 12.2 (2016), p. 20150623.

[BEK21] N. Biswas, A. Etheridge, and A. Klimek. “The spatial Lambda-Fleming-Viot process with fluctuating selection”. In: *Electronic Journal of Probability* 26 (2021), pp. 1–51.

[BEV10] N. Barton, A. Etheridge, and A. Véber. “A new model for evolution in a spatial continuum”. In: *Electronic Journal of Probability* 15 (2010), pp. 162–216.

[BG97] L. Bertini and G. Giacomin. “Stochastic Burgers and KPZ equations from particle systems”. In: *Communications in Mathematical Physics* 183.3 (1997), pp. 571–607.

[BK20] J. Blath and N. Kurt. “Population genetic models of dormancy”. In: *arXiv preprint arXiv:2012.00810* (2020).

[BK93] J. van den Berg and H. Kesten. “Inequalities for the time constant in first-passage percolation”. In: *Annals of Applied Probability* (1993), pp. 56–80.

[Bla+13] J. Blath et al. “The ancestral process of long-range seed bank models”. In: *Journal of Applied Probability* 50.3 (2013), pp. 741–759.

[Bla+15a] É. Blanchet et al. “Multivariate analysis of polyploid data reveals the role of railways in the spread of the invasive South African Ragwort (Senecio inaequidens)”. In: *Conservation Genetics* 16.3 (2015), pp. 523–533.

[Bla+15b] J. Blath et al. “Genealogy of a Wright-Fisher Model with strong seed-bank component”. In: *XI Symposium on Probability and Stochastic Processes*. Springer. 2015, pp. 81–100.

[Bla+16] J. Blath et al. “A new coalescent for seed-bank models”. In: *Annals of Applied Probability* 26.2 (2016), pp. 857–891.

[Bla+20] J. Blath et al. “Statistical tools for seed bank detection”. In: *Theoretical Population Biology* 132 (2020), pp. 1–15.

[BNss] H. Berestycki and G. Nadin. “Asymptotic spreading for general heterogeneous Fisher-KPP type equations”. In: *Memoirs of the American Mathematical Society* (In press). URL: https://hal.archives-ouvertes.fr/hal-01171334.

[Boe+21a] F. Boenkost et al. “Haldane’s formula in Cannings models: the case of moderately strong selection”. In: *Journal of Mathematical Biology* 83.6 (2021), pp. 1–31.

[Boe+21b] F. Boenkost et al. “Haldane’s formula in Cannings models: the case of moderately weak selection”. In: *Electronic Journal of Probability* 26 (2021), pp. 1–36.

[Bor+15] B. Borgy et al. “Dynamics of weeds in the soil seed bank: a hidden Markov model to estimate life history traits from standing plant time series”. In: *PLoS one* 10.10 (2015), e0139278. DOI: 10.1371/journal.pone.0139278.

[Bou+12] E. Bouin et al. “Invasion fronts with variable motility: phenotype selection, spatial sorting and wave acceleration”. In: *Comptes Rendus Mathematique* 350.15-16 (2012), pp. 761–766.

[Bra+16] C.J.A. Bradshaw et al. “Massive yet grossly underestimated global costs of invasive insects”. In: *Nature Communications* 7.1 (2016), pp. 1–8.

[Cal+18] V. Calvez et al. “Non-local competition slows down front acceleration during dispersal evolution”. In: *arXiv preprint arXiv:1810.07634* (2018).

[CD16] S. Chatterjee and P.S. Dey. “Multiple phase transitions in long-range first-passage percolation on square lattices”. In: *Communications on Pure and Applied Mathematics* 69.2 (2016), pp. 203–256.

[CD81] T. Cox and R. Durrett. “Some limit theorems for percolation processes with necessary and sufficient conditions”. In: *Annals of Probability* (1981), pp. 583–603.

[Che+08] P.-O. Cheptou et al. “Rapid evolution of seed dispersal in an urban environment in the weed Crepis sancta”. In: *Proceedings of the National Academy of Sciences* 105.10 (2008), pp. 3796–3799.

[CHS19] F. Cordero, S. Hummel, and E. Schertzer. “General selection models: Bernstein duality and minimal ancestral structures”. In: *arXiv preprint arXiv:1903.06731* (2019).

[CK19] J. Chetwynd-Diggle and A. Klimek. “Rare mutations in the spatial Lambda-Fleming-Viot model in a fluctuating environment and SuperBrownian motion”. In: *arXiv preprint arXiv:1901.04374* (2019).

[CM07] N. Champagnat and S. Méléard. “Invasion and adaptive evolution for individual-based spatially structured populations”. In: *Journal of Mathematical Biology* 55.2 (2007), pp. 147–188.

[CMR05] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, New York, 2005.

[CP16] A. Chuang and C.R. Peterson. “Expanding population edges: theories, traits, and trade-offs”. In: *Global Change Biology* 22.2 (2016), pp. 494–512.

[Cur+06] M. Currat et al. “Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans"”. In: *Science* 313.5784 (2006), pp. 172–172.

[CW10] C. Cenik and J. Wakeley. “Pacific salmon and the coalescent effective population size”. In: *PLoS One* 5.9 (2010), e13019.

[Def+19] M. Deforet et al. “Evolution at the edge of expanding populations”. In: *The American Naturalist* 194.3 (2019), pp. 291–305.

[Deh+07a] K. Dehnen-Schmutz et al. “A century of the ornamental plant trade and its impact on invasion success”. In: *Diversity and Distributions* 13.5 (2007), pp. 527–534.

[Deh+07b] K. Dehnen-Schmutz et al. “The horticultural trade and ornamental plant invasions in Britain”. In: *Conservation Biology* 21.1 (2007), pp. 224–231.

[DF16] R. Durrett and W.-T. L. Fan. “Genealogies in expanding populations”. In: *Annals of Applied Probability* 26.6 (2016), pp. 3456–3490.

[Dia+21] C. Diagne et al. “High and rising economic costs of biological invasions worldwide”. In: *Nature* 592.7855 (2021), pp. 571–576.

[Dit+13] D. van Ditmarsch et al. “Convergent evolution of hyperswarming leads to impaired biofilm formation in pathogenic bacteria”. In: *Cell Reports* 4.4 (2013), pp. 697–708.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), pp. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[DPC11] A. Dornier, V. Pons, and P.-O. Cheptou. “Colonization and extinction dynamics of an annual plant metapopulation in an urban environment”. In: *Oikos* 120.8 (2011), pp. 1240–1246.

[DSV03] R.L.H. Dennis, T.G. Shreeve, and H. Van Dyck. “Towards a functional resource-based concept for habitat: a butterfly biology viewpoint”. In: *Oikos* (2003), pp. 417–426.

[DT95] P. Donnelly and S. Tavare. “Coalescents and genealogical structure under neutrality”. In: *Annual Review of Genetics* 29.1 (1995), pp. 401–421.

[Dur18] R. Durrett. “Genealogies in growing solid tumors”. In: *bioRxiv* (2018), p. 244160.

[Dur84] R. Durrett. “Oriented percolation in two dimensions”. In: *The Annals of Probability* 12.4 (1984), pp. 999–1040.

[Ede61] M. Eden. “A two-dimensional growth process”. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 4. University of California Press Berkeley. 1961, pp. 223–239.

[EFP09] L. Excoffier, M. Foll, and R.J. Petit. “Genetic consequences of range expansions”. In: *Annual Review of Ecology, Evolution, and Systematics* 40 (2009), pp. 481–501.

[EFP17] A. Etheridge, N. Freeman, and S. Penington. “Branching Brownian motion, mean curvature flow and the motion of hybrid zones”. In: *Electronic Journal of Probability* 22 (2017), pp. 1–40.

[EFS17] A. Etheridge, N. Freeman, and D. Straulino. “The Brownian net and selection in the spatial Lambda-Fleming-Viot process”. In: *Electronic Journal of Probability* 22 (2017).

[EK19] A. Etheridge and T.G. Kurtz. “Genealogical constructions of population models”. In: *Annals of Probability* 47.4 (2019), pp. 1827–1910.

[EK86] S.N. Ethier and T.G. Kurtz. *Markov processes : characterization and convergence*. Wiley, 1986.

[ELC04] C.A. Edmonds, A.S. Lillie, and L.L. Cavalli-Sforza. “Mutations arising in the wave front of an expanding population”. In: *Proceedings of the National Academy of Sciences* 101.4 (2004), pp. 975–979.

[EP20] A.M. Etheridge and S. Penington. “Genealogies in bistable waves”. In: *arXiv preprint arXiv:2009.03841* (2020).

[Eth+17] A. Etheridge et al. “Branching Brownian motion and selection in the spatial Lambda-Fleming–Viot process”. In: *Annals of Applied Probability* 27.5 (2017), pp. 2605–2645.

[Eth08] A.M. Etheridge. “Drift, draft and structure : some mathematical models of evolution”. In: *Banach Center Publications* 80 (2008), pp. 121–144.

[Eth11] A. Etheridge. *Some Mathematical Models from Population Genetics: École D’Été de Probabilités de Saint-Flour XXXIX-2009*. Vol. 2012. Springer Science & Business Media, 2011.

[EV12] A. Etheridge and A. Véber. “The spatial Lambda-Fleming-Viot process on a large torus: Genealogies in the presence of recombination”. In: *Annals of Applied Probability* 22.6 (2012), pp. 2165–2209.

[EVY20] A. Etheridge, A. Véber, and F. Yu. “Rescaling limits of the spatial Lambda-Fleming-Viot process with selection”. In: *Electronic Journal of Probability* 25 (2020), pp. 1–89.

[Fah03] L. Fahrig. “Effects of habitat fragmentation on biodiversity”. In: *Annual review of ecology, evolution, and systematics* 34.1 (2003), pp. 487–515.

[Fan21] W.-T. L. Fan. “Stochastic PDEs on graphs as scaling limits of discrete interacting systems”. In: *Bernoulli* 27.3 (2021), pp. 1899–1941.

[FE20] F. Foutel-Rodier and A. Etheridge. “The spatial Muller’s ratchet: surfing of deleterious mutations during range expansion”. In: *Theoretical Population Biology* 135 (2020), pp. 19–31.

[Fel75] J. Felsenstein. “A pain in the torus: some difficulties with models of isolation by distance”. In: *The American Naturalist* 109.967 (1975), pp. 359–368.

[Fen95] M. Fenner. “Ecology of seed banks”. In: *Seed development and germination* (1995), pp. 507–528.

[Fis37] R. A. Fisher. “The wave of advance of advantageous genes”. In: *Annals of Eugenics* 7.4 (1937), pp. 355–369.

[Fis58] R. A. Fisher. *The genetical theory of natural selection*. Clarenson, Oxford, 1958.

[FP17] R. Forien and S. Penington. “A central limit theorem for the spatial Lambda-Fleming-Viot process with selection”. In: *Electronic Journal of Probability* 22 (2017), pp. 1–68.

[Fré+13] H. Fréville et al. “Inferring seed bank from hidden Markov models: new insights into metapopulation dynamics in plants”. In: *Journal of Ecology* 101.6 (2013), pp. 1572–1580. DOI: 10.1111/1365-2745.12141.

[Fuc+15] Kerissa K Fuccillo et al. “Assessing accuracy in citizen science-based plant phenology monitoring”. In: *International journal of biometeorology* 59.7 (2015), pp. 917–926. DOI: 10.1007/s00484-014-0892-7.

[FW02] R.P. Freckleton and A.R. Watkinson. “Large-scale spatial dynamics of plants: metapopulations, regional ensembles and patchy populations”. In: *Journal of Ecology* 90.3 (2002), pp. 419–434. DOI: 10.1046/j.1365-2745.2002.00692.x.

[Gar+14] O Gargominy et al. “TAXREF v8. 0, référentiel taxonomique pour la France: Méthodologie, mise en oeuvre et diffusion”. In: *Rapport SPN* 42 (2014), p. 2014.

[GHO22] A. Greven, F. den Hollander, and M. Oomen. “Spatial populations with seed-bank: well-posedness, duality and equilibrium”. In: *Electronic Journal of Probability* 27 (2022), pp. 1–88.

[Gon+14] A. González Casanova et al. “Strong seed-bank effects in bacterial evolution”. In: *Journal of Theoretical Biology* 356 (2014), pp. 62–70.

[Got91] N.J. Gotelli. “Metapopulation models: the rescue effect, the propagule rain, and the core-satellite hypothesis”. In: *The American Naturalist* 138.3 (1991), pp. 768–776. DOI: 10.1086/285249.

[Gra+13] E. Graciá et al. “Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion”. In: *Biology Letters* 9.3 (2013), p. 20121091.

[Gri79] D. Griffeath. *Additive and Cancellative Interacting Particle Systems*. Vol. 724. Springer, 1979.

[GS18] A. González Casanova and D. Spanò. “Duality and fixation in Xi-Wright–Fisher processes with frequency-dependent selection”. In: *Annals of Applied Probability* 28.1 (2018), pp. 250–284.

[GS20] A. González Casanova and C. Smadi. “On Lambda-Fleming–Viot processes with general frequency-dependent selection”. In: *Journal of Applied Probability* 57.4 (2020), pp. 1162–1197.

[Hal+07] O. Hallatschek et al. “Genetic drift at expanding frontiers promotes gene segregation”. In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 19926–19930.

[Han04] I. Hanski. “Metapopulation theory, its use and misuse”. In: *Basic and Applied Ecology* 5.3 (2004), pp. 225–229.

[Han11] I. Hanski. “Eco-evolutionary spatial dynamics in the Glanville fritillary butterfly”. In: *Proceedings of the National Academy of Sciences* 108.35 (2011), pp. 14397–14404.

[Har78] T.E. Harris. “Additive set-valued Markov processes and graphical methods”. In: *Annals of Probability* (1978), pp. 355–378.

[HB96] B.C. Husband and S.C.H. Barrett. “A metapopulation perspective in plant population biology”. In: *Journal of Ecology* (1996), pp. 461–469. DOI: 10.2307/2261207.

[Hew00] G. Hewitt. “The genetic legacy of the Quaternary ice ages”. In: *Nature* 405.6789 (2000), pp. 907–913.

[Hew96] G. Hewitt. “Some genetic consequences of ice ages, and their role in divergence and speciation”. In: *Biological journal of the Linnean Society* 58.3 (1996), pp. 247–276.

[HGM97] I.A. Hanski, M.E. Gilpin, and D.E. McCauley. *Metapopulation biology*. Vol. 454. Elsevier, 1997.

[HN08] O. Hallatschek and D.R. Nelson. “Gene surfing in expanding populations”. In: *Theoretical Population Biology* 73.1 (2008), pp. 158–170.

[HN10] O. Hallatschek and D.R. Nelson. “Life at the front of an expanding population”. In: *Evolution: International Journal of Organic Evolution* 64.1 (2010), pp. 193–206.

[HN21] F. den Hollander and S. Nandan. “Spatially inhomogeneous populations with seed-banks: I. Duality, existence and clustering”. In: *Journal of Theoretical Probability* (2021), pp. 1–47.

[HO03] I. Hanski and O. Ovaskainen. “Metapopulation theory for fragmented landscapes”. In: *Theoretical Population Biology* 64.1 (2003), pp. 119–127.

[Hoe63] W. Hoeffding. “Probability Inequalities for Sums of Bounded Random Variables”. In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30. DOI: 10.1080/01621459.1963.10500830.

[HP17] F. den Hollander and G. Pederzani. “Multi-colony Wright-Fisher with seed-bank”. In: *Indagationes Mathematicae* 28.3 (2017), pp. 637–669.

[HS21] I. Hartarsky and R. Szabó. “Generalised oriented site percolation, probabilistic cellular automata and bootstrap percolation”. In: *arXiv preprint arXiv:2103.15621* (2021).

[Hud+15] C.M. Hudson et al. “Virgins in the vanguard: low reproductive frequency in invasion-front cane toads”. In: *Biological Journal of the Linnean Society* 116.4 (2015), pp. 743–747.

[Hue+10] M.A.C. Huergo et al. “Morphology and dynamic scaling analysis of cell colonies with linear growth fronts”. In: *Physical Review E* 82.3 (2010), p. 031903.

[Jar+95] M. Jarry et al. “Modeling the population dynamics of annual plants with seed bank and density dependent effects”. In: *Acta Biotheoretica* 43.1 (1995), pp. 53–65.

[JB85] R. Jullien and R. Botet. “Scaling properties of the surface of the Eden model in d= 2, 3, 4”. In: *Journal of Physics A: Mathematical and general* 18.12 (1985), p. 2279.

[JB87] R. Jullien and R. Botet. “Aggregation and fractal aggregates”. In: *Ann. Telecomm.* 41 (1987), 343–short.

[JMW12] B. Jourdain, S. Méléard, and W.A. Woyczynski. “Lévy flights in evolutionary ecology”. In: *Journal of Mathematical Biology* 65.4 (2012), pp. 677–707.

[Kal06] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[Kaz+21] E. Kazakou et al. “Does seed mass drive interspecies variation in the effect of management practices on weed demography?” In: *Ecology and Evolution* 11.19 (2021), pp. 13166–13174.

[KCE06] S. Klopfstein, M. Currat, and L. Excoffier. “The fate of mutations surfing on the wave of a range expansion”. In: *Molecular Biology and Evolution* 23.3 (2006), pp. 482–490.

[Kee02] M.J. Keeling. “Using individual-based simulations to test the Levins metapopulation paradigm”. In: *Journal of Animal Ecology* (2002), pp. 270–279.

[Kim53] M. Kimura. “’Stepping Stone’ model of population”. In: *Annual Report of the National Institute of Genetics Japan* 3 (1953), pp. 62–63.

[KK03] I. Kaj and S.M. Krone. “The coalescent process in a population with stochastically varying size”. In: *Journal of Applied Probability* 40.1 (2003), pp. 33–48.

[KKL01] I. Kaj, S.M. Krone, and M. Lascoux. “Coalescent theory for seed bank models”. In: *Journal of Applied Probability* (2001), pp. 285–300.

[Kle+08] Michael Kleyer et al. “The LEDA Traitbase: a database of life-history traits of the Northwest European flora”. In: *Journal of Ecology* 96.6 (2008), pp. 1266–1274. DOI: 10.1111/j.1365-2745.2008.01430.x.

[KN97] S.M. Krone and C. Neuhauser. “Ancestral processes with selection”. In: *Theoretical Population Biology* 51.3 (1997), pp. 210–237.

[Koo+17] B. Koopmann et al. “Fisher-Wright model with deterministic seed bank and selection”. In: *Theoretical Population Biology* 114 (2017), pp. 29–39.

[Kor+10] K.S. Korolev et al. “Genetic demixing and evolution in linear stepping stone models”. In: *Reviews of Modern Physics* 82.2 (2010), p. 1691.

[Kow95] I. Kowarik. “Time lags in biological invasions with regard to the success and failure of alien species”. In: *Plant invasions: general aspects and special problems* (1995), pp. 15–38.

[KPP37] A.N. Kolmogorov, I.G. Petrovskii, and N.S. Piskunov. “Étude de l’équation de la chaleur, de la matière et son application à un problème biologique”. In: *Bull. Moskov. Gos. Univ. Mat. Mekh* 1 (1937), p. 125.

[KPZ86] M. Kardar, G. Parisi, and Y.-C. Zhang. “Dynamic scaling of growing interfaces”. In: *Physical Review Letters* 56.9 (1986), p. 889.

[KR20] A. Klimek and T.C. Rosati. “The spatial Lambda-Fleming-Viot process in a random environment”. In: *arXiv preprint arXiv:2004.05931* (2020).

[Kum+15] S. Kumschick et al. “Ecological impacts of alien species: quantification, scope, caveats, and recommendations”. In: *BioScience* 65.1 (2015), pp. 55–63.

[KW64] M. Kimura and G.H. Weiss. “The stepping stone model of population structure and the decrease of genetic correlation with distance”. In: *Genetics* 49.4 (1964), p. 561.

[Lat73] B.D.H. Latter. “The island model of population differentiation: a general solution”. In: *Genetics* 73.1 (1973), pp. 147–157.

[LCP19] S. Le Coz, P.-O. Cheptou, and N. Peyrard. “A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy”. In: *Theoretical Population Biology* 127 (2019), pp. 120–132.

[Len+15] U. Lenz et al. “Looking down in the ancestral selection graph: A probabilistic approach to the common ancestor type distribution”. In: *Theoretical Population Biology* 103 (2015), pp. 27–37.

[Len+21] J. T. Lennon et al. “Principles of seed banks: complexity emerging from dormancy”. In: *Nature Communications* 12.1 (2021), pp. 1–16.

[Lev69] R. Levins. “Some demographic and genetic consequences of environmental heterogeneity for biological control”. In: *American Entomologist* 15.3 (1969), pp. 237–240. DOI: 10.1093/besa/15.3.237.

[Lig85] T.M. Liggett. “An improved subadditive ergodic theorem”. In: *Annals of Probability* 13.4 (1985), pp. 1279–1285.

[LM15] A. Lambert and C. Ma. “The coalescent in peripatric metapopulations”. In: *Journal of Applied Probability* 52.2 (2015), pp. 538–557.

[Lou+21] A. Louvet et al. “Detecting seed bank influence on plant metapopulation dynamics”. In: *Methods in Ecology and Evolution* 12.4 (2021), pp. 655–664.

[Lou21] A. Louvet. “The k-parent spatial Lambda-Fleming-Viot process as a stochastic measure-valued model for an expanding population”. In: *arXiv preprint arXiv:2103.02902* (2021).

[Lou22] A. Louvet. “Extinction threshold and scaling limit of a plant metapopulation model with recurrent extinction events and a seed bank component”. In: *Theoretical Population Biology* 145 (2022), pp. 22–37.

[LPR18] F. Laroche, H. Paltto, and T. Ranius. “Abundance-based detectability in a spatially-explicit metapopulation: a case study on a vulnerable beetle species in hollow trees”. In: *Oecologia* 188.3 (2018), pp. 671–682.

[LV22] A. Louvet and A. Veber. “Growth properties of the infinite-parent spatial Lambda-Fleming Viot process”. In: *arXiv preprint arXiv:2205.03937* (2022).

[Mac+96] N. Machon et al. “Evidence of genetic drift in chestnut populations”. In: *Canadian Journal of Forest Research* 26.5 (1996), pp. 905–908.

[Mac20] N. Machon. “Floristic monitoring of the 1,324 alignment tree bases of 15 streets in the district of Bercy, Paris, France, from 2009 to 2018 (dataset)”. In: *Zenodo* (2020).

[Mar70] T. Maruyama. “Effective number of alleles in a subdivided population”. In: *Theoretical Population Biology* 1.3 (1970), pp. 273–306.

[Mau10] N. Maurel. “De l’introduction à l’invasion: les plantes exotiques en milieu urbain”. PhD thesis. Paris, Muséum National d’Histoire Naturelle, 2010.

[MO22] MNHN and OFB. *Fiche de Veronica persica Poir., 1808. Inventaire National du Patrimoine Naturel (INPN)*. `https://inpn.mnhn.fr/espece/cd_nom/128956`. Accessed: January 6th, 2022. 2003-2022.

[Moi02] Atte Moilanen. “Implications of empirical data quality to metapopulation model parameter estimation and application”. In: *Oikos* 96.3 (2002), pp. 516–530.

[Moi04] A. Moilanen. “SPOMSIM: software for stochastic patch occupancy models of metapopulation dynamics”. In: *Ecological Modelling* 179.4 (2004), pp. 533–550.

[Moi99] A. Moilanen. “Patch occupancy models of metapopulation dynamics: efficient parameter estimation using implicit statistical inference”. In: *Ecology* 80.3 (1999), pp. 1031–1043.

[Mor58] P.A.P. Moran. “Random processes in genetics”. In: *Mathematical proceedings of the Cambridge philosophical society*. Vol. 54. 1. Cambridge University Press. 1958, pp. 60–71.

[MP90] M.J. McDonnell and S.T.A Pickett. “Ecosystem structure and function along urban-rural gradients: an unexploited opportunity for ecology”. In: *Ecology* (1990), pp. 1232–1237.

[MPS00] U. Maurer, T. Peschel, and S. Schmitz. “The flora of selected urban land-use types in Berlin and Potsdam with regard to nature conservation in cities”. In: *Landscape and Urban Planning* 46.4 (2000), pp. 209–215.

[MS95] C. Mueller and R.B. Sowers. “Random travelling waves for the KPP equation with noise”. In: *Journal of Functional Analysis* 128.2 (1995), pp. 439–498.

[MSH98] A. Moilanen, A.T. Smith, and I. Hanski. “Long-term dynamics in a metapopulation of the American pika”. In: *The American Naturalist* 152.4 (1998), pp. 530–542.

[Mur+07] A. Muratet et al. “The role of urban structures in the distribution of wasteland flora in the greater Paris area, France”. In: *Ecosystems* 10.4 (2007), pp. 661–671.

[MW67] R.H. MacArthur and E.O. Wilson. *The theory of island biogeography*. Princeton University Press, 1967.

[NK97] C. Neuhauser and S.M. Krone. “The genealogy of samples in models with selection”. In: *Genetics* 145.2 (1997), pp. 519–534.

[Not90] M. Notohara. “The coalescent and the genealogical process in geographically structured population”. In: *Journal of Mathematical Biology* 29.1 (1990), pp. 59–75.

[Not97] M. Notohara. “The number of segregating sites in a sample of DNA sequences from a geographically structured population”. In: *Journal of Mathematical Biology* 36.2 (1997), pp. 188–200.

[OH04] O. Ovaskainen and I. Hanski. “From individual behavior to metapopulation dynamics: unifying the patchy population and classic metapopulation models”. In: *The American Naturalist* 164.3 (2004), pp. 364–377.

[Oma+18] M. Omar et al. “Drivers of the distribution of spontaneous plant communities and species within urban tree bases”. In: *Urban Forestry & Urban Greening* 35 (2018), pp. 174–191. DOI: 10.1016/j.ufug.2018.08.018.

[Oma+19] M. Omar et al. “Colonization and extinction dynamics among the plant species at tree bases in Paris (France)”. In: *Ecology and Evolution* 9.15 (2019), pp. 8414–8428.

[Ozg+06] A. Ozgul et al. “Spatiotemporal variation in survival rates: implications for population dynamics of yellow-bellied marmots”. In: *Ecology* 87.4 (2006), pp. 1027–1037.

[Pai+16] D.R. Paini et al. “Global threat to agriculture from invasive species”. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7575–7579.

[PE15] S. Peischl and L. Excoffier. “Expansion load: recessive mutations and the role of standing genetic variation”. In: *Molecular Ecology* 24.9 (2015), pp. 2084–2094.

[Pei+13] S. Peischl et al. “On the accumulation of deleterious mutations during range expansions”. In: *Molecular Ecology* 22.24 (2013), pp. 5972–5982.

[Pit99] J. Pitman. “Coalescents with multiple collisions”. In: *Annals of Probability* (1999), pp. 1870–1902.

[Plu+18] M. Pluntz et al. “A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora”. In: *Ecology Letters* 21.9 (2018), pp. 1311–1318.

[Pyš98] P. Pyšek. “Alien and native species in Central European urban floras: a quantitative comparison”. In: *Journal of Biogeography* 25.1 (1998), pp. 155–163.

[Rab89] L.R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286. DOI: 10.1109/5.18626.

[Ram+05] S. Ramachandran et al. “Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa”. In: *Proceedings of the National Academy of Sciences* 102.44 (2005), pp. 15942–15947.

[Rat+16] F.L.W. Ratnieks et al. “Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers”. In: *Methods in Ecology and Evolution* 7.10 (2016), pp. 1226–1235. DOI: 10.1111/2041-210X.12581.

[Ric73] D. Richardson. “Random growth in a tessellation”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 74. 3. Cambridge University Press. 1973, pp. 515–528.

[Rob18] S. Robin. *Models with Hidden Structure with Applications in Biology and Genomics*. Jan. 2018.

[Ros+02] N.A. Rosenberg et al. “Genetic structure of human populations”. In: *Science* 298.5602 (2002), pp. 2381–2385.

[Ros81] H. Rost. “Non-equilibrium behaviour of a many particle process: Density profile and local equilibria”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 58.1 (1981), pp. 41–53.

[Sak+01] A.K. Sakai et al. “The population biology of invasive species”. In: *Annual Review of Ecology and Systematics* 32.1 (2001), pp. 305–332.

[Sch+07] K.B. Schroeder et al. “A private allele ubiquitous in the Americas”. In: *Biology Letters* 3.2 (2007), pp. 218–223.

[Sch00] J. Schweinsberg. “A Necessary and Sufficient Condition for the Lambda-Coalescent to Come Down from Infinity.” In: *Electronic Communications in Probability* 5 (2000), pp. 1–11.

[Sel+20] T. Sellinger et al. “Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data”. In: *PLoS Genetics* 16.4 (2020), e1008698.

[Sep09] T. Seppäläinen. “Lecture notes on the corner growth model”. In: *Unpublished notes* (2009).

[SH91] M. Slatkin and R.R. Hudson. “Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations.” In: *Genetics* 129.2 (1991), pp. 555–562.

[She+11] D.S. Shepard et al. “Economic impact of dengue illness in the Americas”. In: *The American Journal of Tropical Medicine and Hygiene* 84.2 (2011), p. 200.

[Shp+10] M. Shpak et al. “A structured coalescent process for seasonally fluctuating populations”. In: *Evolution: International Journal of Organic Evolution* 64.5 (2010), pp. 1395–1409.

[Ska13] M. Skala. “Hypergeometric tail inequalities: ending the insanity”. In: *arXiv preprint arXiv:1311.5939* (2013).

[Sla77] M. Slatkin. “Gene flow and genetic drift in a species subject to frequent local extinctions”. In: *Theoretical Population Biology* 12.3 (1977), pp. 253–262.

[Smi+06] R.M. Smith et al. “Urban domestic gardens (IX): composition and richness of the vascular plant flora, and implications for native biodiversity”. In: *Biological Conservation* 129.3 (2006), pp. 312–322.

[SNS08] J.-C. Svenning, S. Normand, and F. Skov. “Postglacial dispersal limitation of widespread forest plant species in nemoral Europe”. In: *Ecography* 31.3 (2008), pp. 316–326.

[SNS17] M.A. Stoffel, S. Nakagawa, and H. Schielzeth. “rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models”. In: *Methods in Ecology and Evolution* 8.11 (2017), pp. 1639–1644.

[Sot+15] A. Sottoriva et al. “A Big Bang model of human colorectal tumor growth”. In: *Nature Genetics* 47.3 (2015), pp. 209–216.

[SS07] J.-C. Svenning and F. Skov. “Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation?” In: *Ecology Letters* 10.6 (2007), pp. 453–460.

[SSI04] A. Sano, A. Shimizu, and M. Iizuka. “Coalescent process with fluctuating population size and its effective size”. In: *Theoretical Population Biology* 65.1 (2004), pp. 39–48.

[Suk04] H. Sukopp. “Human-caused impact on preserved vegetation”. In: *Landscape and Urban Planning* 68.4 (2004), pp. 347–355. DOI: 10.1016/S0169-2046(03)00152-X.

[SW78] R.T. Smythe and J.C. Wierman. *First-passage percolation on the square lattice*. Vol. 671. Springer, 1978.

[Tel+11] A. Tellier et al. “Inference of seed bank parameters in two wild tomato species using ecological and genetic data”. In: *Proceedings of the National Academy of Sciences* 108.41 (2011), pp. 17052–17057.

[Tem02] A. Templeton. “Out of Africa again and again”. In: *Nature* 416.6876 (2002), pp. 45–51.

[Tra+07] J.M.J. Travis et al. “Deleterious mutations can surf to high densities on the wave front of an expanding population”. In: *Molecular Biology and Evolution* 24.10 (2007), pp. 2334–2343.

[TV09] J. Taylor and A. Véber. “Coalescent processes in subdivided populations subject to recurrent mass extinctions”. In: *Electronic Journal of Probability* 14 (2009), pp. 242–288.

[Van+07] S. Van der Veken et al. “Over the (range) edge: a 45-year transplant experiment with the perennial forest herb Hyacinthoides non-scripta”. In: *Journal of Ecology* 95.2 (2007), pp. 343–351.

[VK07] M. Von der Lippe and I. Kowarik. “Long-distance dispersal of plants by vehicles as a driver of plant invasions”. In: *Conservation Biology* 21.4 (2007), pp. 986–996. DOI: 10.1111/j.1523-1739.2007.00722.x.

[VW15] A. Veber and A. Wakolbinger. “The spatial Lambda-Fleming-Viot process: An event-based construction and a lookdown representation”. In: 51.2 (2015), pp. 570–598.

[WA01] J. Wakeley and N. Aliacar. “Gene genealogies in a metapopulation”. In: *Genetics* 159.2 (2001), pp. 893–905.

[Wak04] J. Wakeley. “Metapopulation models for historical inference”. In: *Molecular Ecology* 13.4 (2004), pp. 865–875.

[Wak98] J. Wakeley. “Segregating sites in Wright’s island model”. In: *Theoretical Population Biology* 53.2 (1998), pp. 166–174.

[WCV16] J.R. Walsh, S.R. Carpenter, and M.J. Vander Zanden. “Invasive species triggers a massive loss of ecosystem services through a trophic cascade”. In: *Proceedings of the National Academy of Sciences* 113.15 (2016), pp. 4081–4085.

[WG04] E. Weber and D. Gut. “Assessing the risk of potentially invasive plant species in central Europe”. In: *Journal for Nature Conservation* 12.3 (2004), pp. 171–179.

[Wil+93] M.A.J. Williams et al. “Quaternary environments”. In: *Quaternary environments*. Edward Arnold Publishers Ltd, 1993.

[Wil98] H.M. Wilkinson-Herbots. “Genealogy and subpopulation differentiation under various models of population structure”. In: *Journal of Mathematical Biology* 37.6 (1998), pp. 535–585.

[WLB95] C.Y. Wang, P.L. Liu, and J.B. Bassingthwaighte. “Off-lattice Eden-C cluster growth model”. In: *Journal of Physics A: Mathematical and general* 28.8 (1995), p. 2141.

[WM90] M.C. Whitlock and D.E. McCauley. “Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups”. In: *Evolution* 44.7 (1990), pp. 1717–1724.

[Wri31] S. Wright. “Evolution in Mendelian populations”. In: *Genetics* 16.2 (1931), p. 97.

[ŽT12] D. Živković and A. Tellier. “Germ banks affect the inference of past demographic events”. In: *Molecular Ecology* 21.22 (2012), pp. 5434–5446.

Contents

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre: Modèles probabilistes de génétique des populations pour les populations en expansion

Mots clés: modèles de populations en expansion, processus stochastiques, génétique des populations, théorèmes limites, processus à valeurs mesures, banques de graines

Résumé: Cette thèse porte sur la construction et l'étude de modèles probabilistes de génétique des populations pour les populations en expansion. Nous construisons ces modèles à partir d'un concept issu de la théorie des systèmes de particules en interaction, qui consiste à représenter les sites vides comme occupés par des particules d'un type spécifique. Ces "individus fantômes" nous permettent de maintenir artificiellement un nombre d'individus constant, et de construire des processus duaux encodant les généalogies. Dans nos modèles, les individus fantômes peuvent aussi se reproduire, modélisant ainsi les fluctuations stochastiques du nombre d'individus, mais avec un très fort désavantage sélectif face aux individus "réels". Nous appliquons d'abord le concept d'individus fantômes à un processus à valeurs mesure qui décrit la dynamique de reproduction d'une population vivant dans un espace continu. Nous construisons la limite de ce processus lorsque la "sélection" contre les individus fantômes devient infiniment forte. Le processus limite semble être un équivalent du modèle d'Eden en espace continu. Nous étudions la dynamique d'expansion des individus réels dans le processus limite, et montrons que la croissance de la région qu'ils occupent est linéaire en temps. Nous nous intéressons ensuite à une variante du modèle de Wright-Fisher structuré spatialement, incluant une banque de graines et des extinctions locales fréquentes. Ceci est motivé par une question d'intérêt en écologie : comprendre la dynamique des plantes dans les pieds d'arbres d'alignement en ville. Dans une étude préliminaire sur un jeu de données réelles, nous montrons qu'il est nécessaire de prendre en compte la présence potentielle d'une banque de graines pour répondre à cette question. Nous utilisons notre variante du modèle de Wright-Fisher pour montrer l'existence d'une probabilité critique d'extinction de patch dépendant des paramètres de banque de graines au delà de laquelle une expansion de population n'est pas possible. Nous étudions la limite de ce processus dans un régime de sélection forte, et montrons qu'il converge vers un modèle de présence/absence. Ce modèle limite appartient à une famille de modèles très utilisés en écologie des métapopulations.

Title: Probabilistic population genetics models for expanding populations

Keywords: models for expanding populations, stochastic processes, population genetics, limit theorems, measure-valued processes, seed banks

Abstract: This thesis focuses on the construction and study of stochastic population genetics models for expanding populations. We build different models using a concept from the theory of interacting particles systems, where empty sites are represented as occupied by particles with a specific type. These "ghost individuals" allow us to artificially keep population sizes constant, and build dual processes encoding genealogies. In our models, ghost individuals can reproduce as well in order to account for stochastic fluctuations in population sizes, but with a very strong selective disadvantage against "real" individuals. We first apply the concept of ghost individuals to a measure-valued process describing the reproduction dynamics of a population living in a continuous space. We construct the limit of the process when "selection" against ghost individuals becomes infinitely strong. The limiting process seems to be a space continuous equivalent of the Eden growth model. We study the expansion dynamics of real individuals in the limiting process, and show that the growth of the region they occupy is linear in time. We then focus on a variant of the spatially-structured Wright-Fisher model with a seed bank component and featuring frequent local extinction events. This was motivated by a question of ecological interest: understanding plant dynamics in urban tree bases. In a preliminary study on a real dataset, we show that it is necessary to account for the potential presence of a seed bank in order to answer this question. We use our variant of the Wright-Fisher model to show the existence of a critical patch extinction probability depending on seed bank parameters, above which a population expansion is not possible. We study the limit of the process under a strong selection regime, and show that it corresponds to an occupancy-based model. This limiting process belongs to a family of models widely used in metapopulation ecology.

Institut Polytechnique de Paris
91120 Palaiseau, France