Modélisation de la dynamique des adventices dans un agroécosystème

Sebastian Le Coz

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Sebastian LE COZ

Le 12 mars 2019

Modélisation de la dynamique des adventices dans un agroécosystème.

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Mathématiques et Applications**

Unité de recherche :
MIAT-INRA : Unité de Mathématiques et Informatique Appliquées Toulouse

Thèse dirigée par
Nathalie PEYRARD et Pierre-Olivier CHEPTOU

Jury

M. Frédéric GOSSELIN, Rapporteur
M. Nikolaos LIMNIOS, Rapporteur
M. Jean-yves TOURNERET, Examinateur
M. Xavier REBOUD, Examinateur
Mme Nathalie PEYRARD, Directrice de thèse
M. Pierre-Oliver CHEPTOU, Co-directeur de thèse

Résumé

De nombreuses espèces ont un stade dormant dans leur cycle de vie, comme les graines chez les plantes. Ces espèces ont recours à plusieurs méthodes afin de survivre dans l'environnement. En particulier, les plantes sont connues pour avoir une stratégie de survie dépendant de la dormance et de la dispersion des graines. Le modèle de métapopulation est souvent utilisé afin d'étudier la dynamique régionale d'espèces. Cependant, celui-ci ne modélisant pas de stade dormant dans la dynamique de l'espèce, appliqué à une espèce avec stade dormant, il peut amener à prédire l'extinction de l'espèce au sein d'un patch alors que celle-ci est présente sous forme dormante. Du fait que le stade dormant d'une espèce soit difficilement observable en pratique, si l'on veut inclure la dormance dans un modèle, il est préférable d'utiliser des variables cachées pour modéliser ce stade. Plusieurs modèles avec structure Markovienne et variables cachées ont déjà été utilisés pour étudier les espèces avec stade caché. Cependant ils présentent tous des limites : la modélisation des données en présence/absence, le stade dormant limité à une année ainsi que la colonisation entre patchs qui n'est pas prise en compte. Je propose ici un modèle de chaîne de Markov cachée multidimensionnelle avec retour des données qui permet de décrire la dynamique d'espèces avec stade caché où seuls les stades observables sont à l'origine d'interactions entre patchs. Ces interactions entre patchs sont modélisées à partir de l'influence indistinguable des populations observables des patchs voisins sur une population observable ou cachée. Ce modèle, utilisant des données en classes d'abondance, permet une dormance potentiellement infinie. J'ai montré que la complexité algorithmique de l'estimation des paramètres du modèle n'est pas exponentielle, comme on pourrait s'y attendre, mais seulement linéaire en le nombre de patchs. Les résultats sur simulations montrent qu'il est possible de restaurer l'état d'une population en stade caché ainsi que de prédire le prochain état d'une population observable. Les résultats sur données de plantes adventices mettent en évidence la survie de banque de graines comme le processus ayant le plus d'influence sur l'état de la banque de graines. Le modèle permet d'étudier de façon efficace la dynamique de plantes adventices ainsi que d'autres espèces avec stade caché.

Abstract

Many species have a dormant stage in their life cycle, such as seeds for plants. Theses species have different types of survival strategies. In particular, plants are known have survival strategies dependent on dormancy and dispersal of seeds. The metapopulation model, which does not consider a dormancy stage and is often used to analyse a species' dynamic, applied to a species which undergoes dormancy can lead to wrongly declare extinction in a patch where dormant individuals can still be present. In order to include dormancy in a model it is preferable to use hidden variables to model dormant individuals as they are often unobservable. Several Markovian models with hidden variables have already been proposed to study species with hidden stages. However, they all have different limitations : only presence/absence observations are modelled ; the dormancy stage is limited to one year or colonisation from neighbour patches is not taken into account. We propose a hidden Markov model with data feedback which describes the local and regional dynamics of a species with hidden stages where only observables stages may influence other patchs. The model allows species to undergo potentially time infinite dormancy using abundance classes. One would expect estimation, restoration and prediction of the next non-dormant populations to have an exponential computational time in terms of patches, however we have demonstrated that estimation, restoration and prediction are all achievable in a linear in terms of patches. The regional dynamic is modeled using the indistinguishable influence of neighbour non-dormant populations states on a dormant or non dormant population. Numerical experiments on simulated data show that the state dormant populations can easily be retrieved as well as the future non-dormant populations' state. Results on weed species highlight that the state of the seed bank is mostly influenced by seed survival. Our framework provides a simple and efficient tool that could be further exploited to analyse and compare annual plants' dynamics, like weeds survival strategies in crop fields and even for species with hidden stages.

Remerciement

Je remercie tous mes collègues de l'INRA pour leur bonne humeur et la bonne ambiance au sein de l'unité et plus particulièrement mes collègues de bureau. Je remercie toutes les personnes qui ont été présentes et m'ont soutenu tout au long de ma thèse. Mes parents, qui ont toujours cru en moi pendant mes études, m'ont accompagné pendant ces trois années et je les en remercie. Je remercie ma famille pour leur soutien. Je remercie aussi Véronique Gauthier et Jade Nardi pour m'avoir aidé à rendre mon document de thèse plus clair en corrigeant la plupart les fautes d'orthographe. Je voudrais remercier tout particulièrement Jade Nardi pour m'avoir supporté au quotidien ainsi que pour toutes ses remarques et critiques constructives, sur mon document de thèse. Je remercie Stéphane Cordeau pour les données d'Époisses.

Je remercie Frédéric Gosselin ainsi que Nikolaos Limnios pour avoir accepté d'être rapporteurs pour ma thèse. Leurs commentaires ont été d'une grande clarté et je leur en remercie. J'aimerais remercier mes deux directeurs de thèse pour leur soutien et leur présence constante durant ces trois ans. Je remercie tous les autres personnes que je n'ai pas mentionnées. J'aimerais finir cette section en remerciant la région Occitanie, l'ANR AGROBIOS ainsi que l'INRA pour avoir financé cette thèse. Grâce à toutes les personnes, citées j'ai pris un grand plaisir à faire de la recherche dans ce domaine.

Table des matières

Chapitre 1

Introduction

Les adventices sont des espèces de plantes dans des champs agricoles qui sont en compétition avec la culture semée. Leur présence est souvent la cause de nuisances économiques même si elles engendrent parfois une augmentation du rendement de la culture semée car elles contribuent au maintien des abeilles sauvages ou domestiques, utiles pour la pollinisation des cultures (Bretagnolle and Gaba, 2015). Ainsi, plusieurs méthodes de gestion, rotation de culture ou de type désherbage ont été mises en place afin de limiter leur impact sur le rendement de la culture semée. Par ailleurs, ces dernières décennies, les enjeux sociétaux et environnementaux ont conduit les agriculteurs à repenser leurs méthodes de gestion. En effet, les méthodes chimiques, ayant des effets néfastes sur la biodiversité ainsi que sur la santé, rendent les bio-agresseurs résistants si elles sont utilisées fréquemment (Green, 2009). D'un point de vue écologique et environnemental, il est préférable de recourir à des méthodes de régulations biologiques pour maintenir une production agricole. Cependant, cela nécessite une bonne compréhension des facteurs biotiques et abiotiques qui influencent la survie des adventices. Au-delà de la gestion, comprendre le fonctionnement des adventices peut être utile dans une optique de conservation de la flore adventice. Ma thèse s'inscrit dans ce contexte agroécologique avec pour objectif d'apporter des outils mathématiques qui aideront à identifier les facteurs les plus influents sur la survie d'une population d'adventices, tels que le type de culture semée ou bien le type de désherbage utilisé.

Afin de pouvoir déterminer les facteurs les plus influents, il est important de connaître les processus susceptibles d'influencer la survie d'une population. Ces facteurs, environnementaux ou humains, sont nombreux et varient selon les espèces. En outre, ces facteurs peuvent aussi bien avoir un effet sur la dynamique locale d'une espèce que sur la dynamique régionale. La dynamique régionale d'une espèce est liée à sa capacité à coloniser des champs voisins. Elle dépend de facteurs comme le vent, les rivières, les animaux et les outils agricoles. De plus, la capacité à coloniser varie selon les espèces. La persistance locale des adventices dépend de la persistance des populations de graines dans le sol et des plantes face à leur environnement ainsi que des stratégies de gestion mises en place au sein du champ. D'après Chadoeuf-Hannel (1985), la densité de graines dans un sol agricole peut varier entre 0 et plusieurs millions de graines par mètre carré. Généralement, entre 70 et 90% de ces graines proviennent de quelques espèces d'adventices dominantes qui résistent bien aux méthodes de gestion (Wilson and Worsham, 1988). Le risque d'extinction dans un champ est réduit par la dormance potentielle des graines. La durée de dormance s'étend de un an à plusieurs décennies selon les espèces et l'environnement. A cause de tous ces facteurs, il est difficile de savoir si l'apparition d'une plante adventice dans un champ résulte de la germination d'une graine issue de la colonisation ou d'une graine locale sortie de son état de dormance pour germer.

Modéliser et estimer la dynamique des adventices répondrait à plusieurs questions que les écologues, les agronomes et les agriculteurs se posent. L'influence des tracteurs est-elle négligeable

dans la dispersion des graines adventices ? Existe-t-il un compris entre colonisation et dormance ? Ce compromis dépend-il de la culture en place ? Un modèle pourrait servir à identifier la cause de l'apparition d'une adventice dans un champ, à l'aide de prédictions de l'état de la banque de graines. Il permettrait également d'optimiser les stratégies de gestion des adventices soit pour réduire le recours aux herbicides soit pour aider un agriculteur dans la prise de décision de la rotation de culture.

En écologie, un patch est une unité spatiale relativement homogène qui diffère de ses voisins. L'un des modèles les plus répandus pour étudier la distribution d'une population à travers plusieurs patchs est le modèle de métapopulation développé par Levins et al. (1969), qui utilise un paramètre de colonisation et un paramètre d'extinction. Le modèle de Levins considère que la colonisation est dépendante de la fraction de patchs occupés. Plusieurs études ont utilisé le modèle de métapopulation sur différents types d'espèces comme les papillons (Hanski and Thomas, 1994) ou les scarabées (Cornelisse et al., 2013). Cependant, pour les espèces ayant un état de dormance, les modèles de métapopulation ne semblent pas pertinents (Freckleton and Watkinson, 2002 ; Bullock et al., 2006). En effet, quand un modèle de métapopulation est utilisé sur la population non dormante de l'espèce ayant un état de dormance, l'extinction locale peut être déclarée alors que l'espèce est toujours présente sous forme dormante. Cela implique que, pour un modèle de métapopulation, la présence d'une population non dormante après une période d'absence est le résultat d'une colonisation. Cependant, l'apparition soudaine d'une population non dormante peut être due à une population sortie de dormance. Ainsi, Fréville et al. (2013) ont montré que les modèles de métapopulation ont tendance à surestimer les paramètres de colonisation et d'extinction en présence de banque de graines.

Plusieurs modèles ont été créés afin de représenter la dynamique d'une espèce ayant un stade de dormance à l'aide d'informations sur les populations dormantes et non-dormantes (Cohen, 1966 ; Levin et al., 1984 ; Jarry et al., 1995 ; Amarasekare and Possingham, 2001 ; Mistro et al., 2005 ; Han et al., 2014). Cependant, la collecte de données est souvent faite uniquement sur les populations non-dormantes car les populations dormantes sont plus difficilement observables. Cela implique que pour estimer la dynamique locale et régionale des adventices, il serait préférable d'utiliser un modèle dans lequel la banque de graines est modélisée par une variable cachée. La dynamique locale des espèces avec stade caché a déjà été étudiée avec observation incomplète (David et al., 2010 ; Quintana-Ascencio et al., 2011 ; Lamy et al., 2013 ; Fréville et al., 2013 ; Borgy et al., 2015 ; Manna et al., 2017). Les modèles Lamy et al. (2013) ; Fréville et al. (2013) ; Borgy et al. (2015) ; Manna et al. (2017) utilisent une structure markovienne : l'état des populations à un temps donné ne dépend que de l'état des populations au temps précédent. Les chaînes de Markov cachées représentent une des extensions classiques aux chaînes de Markov. On appelle chaîne de Markov cachée un couple formé d'une chaîne de Markov dont les états ne sont pas visibles et d'une suite d'observations qui dépendent chacune de de l'état actuel de la chaîne de Markov. De plus, les observations n'influencent pas la chaîne de Markov cachée. Cette hypothèse implique que la dynamique d'une espèce avec stade caché ne peut être directement modélisée à l'aide d'une chaîne de Markov cachée. En effet, en considérant les populations de graines dans le sol comme les variables cachées et les populations de plantes comme les variables observables, les populations observables devraient influencer les populations cachées. Dans un contexte d'adventices, les plantes (les populations observables) produisent de nouvelles graines qui vont alimenter la banque de graines (la population cachée). Borgy et al. (2015) ont étendu les chaînes de Markov cachées pour inclure cette dépendance. On appellera ce modèle Hidden Markov Model with data feedback (HMM-DF).

Le problème avec les modèles David et al. (2010) ; Quintana-Ascencio et al. (2011) ; Lamy et al. (2013) ; Fréville et al. (2013) ; Borgy et al. (2015) réside dans le fait que la dynamique régionale n'est pas modélisée. La contribution de chaque patch dans le processus de colonisation n'est pas prise en compte. La colonisation, quand elle est modélisée, l'est seulement par pluie de graines.

Plusieurs extensions du Hidden Markov Model (HMM) pour inclure la dépendance entre patchs existent. Les Factorial HMM (Ghahramani and Jordan, 1997) ou les Coupled HMM sont des modèles qui étendent les HMM avec dépendance entre patchs. Cependant, les FHMM considèrent que toutes les variables cachées au temps n émettent une observation groupée qui n'influence pas les variables cachées au temps suivant. Cela implique qu'un FHMM n'est pas adéquat pour modéliser le processus de colonisation. Dans un CHMM (Brand et al., 1997), la variable cachée d'une chaîne influence les autres variables cachées des autres chaînes au temps suivant. Si une espèce avec stade caché était modélisée avec un CHMM, cela impliquerait que les populations cachées pourraient coloniser d'autres patchs. Si cela peut être le cas pour certaines espèces, ce n'est pas le cas pour les plantes car leurs graines sont immobiles dans la banque de graines. En plus de ne pas avoir une structure adéquate pour les adventices, l'estimation exacte des FHMM et CHMM est de complexité algorithmique exponentielle en le nombre de chaînes.

L'objectif de cette thèse est de proposer une structure de HMM levant les limites des modèles précédents pour mieux modéliser la dynamique locale et régionale des adventices. Mes contributions portent sur 3 aspects.

Nous établissons un modèle markovien caché multidimensionnel avec retour de données (MHMM-DF) qui inclut la dynamique locale et régionale d'une espèce avec stade caché, la dynamique régionale correspondant à la colonisation de populations observables vers d'autres populations. Le modèle considère que les populations cachées et observables dépendent les unes des autres de façon stochastique. L'utilisation de données de comptage dans un modèle permet des prédictions plus précises que de simples données de type présence/absence. En contrepartie, l'estimation à grande échelle sur une multitude de parcelles peut s'avérer très coûteuse en termes de complexité algorithmique. Par conséquent, on suppose que les populations sont modélisées en classes d'abondance. Le MHMM-DF est entièrement défini par la structure qui représente les interactions entre les populations cachées et observées présentes dans les différents patchs ainsi que par les probabilités de transition. L'état de la population cachée dans un patch au temps n est décrit par le résultat de 4 processus : (i) la survie de la population cachée, (ii) la production locale de nouveaux individus cachés, (iii) la colonisation venant des populations observables voisines et (iv) la colonisation venant de l'extérieur. La colonisation extérieure correspond à la colonisation venant des populations observables de patchs extérieurs à l'étude. L'état de la population observable dans un patch au temps n est décrit par le résultat de 3 processus : (i) la survie de la population observable, (ii) la migration de la population observable et (iii) la production locale de nouveaux individus observables. Le modèle défini ci-dessus permet de modéliser des dynamiques avec dormance plus complexes que celle des adventices. En effet, le sous-modèle correspondant aux adventices n'a pas de processus de migration de populations observables vers d'autres populations observables. Dans le cas où les adventices sont des plantes annuelles, si le pas de temps entre les variables représente une année, alors le modèle considère la survie des populations observables comme inexistante.

La collecte de données se fait rarement pendant de longues périodes, c'est pourquoi les échantillons pour l'estimation sont de petite taille. Pour pallier ce problème, nous proposons 2 versions paramétriques du MHMM-DF qui sont génériquement identifiables. Il en résulte que le MHMM-DF paramétrique possède au maximum 9 paramètres, chaque paramètre étant associé à un processus biologique spécifique.

L'estimation repose sur l'algorithme Expectation Maximisation (EM Dempster et al. (1977)) qui cherche les paramètres maximisant l'espérance du logarithme de la vraisemblance du modèle adventice complet à l'aide du Forward-Backward amélioré. Une estimation naïve des paramètres sur MHMM-DF a une complexité algorithmique exponentielle en le nombre de patchs. Afin de réduire le temps de calcul lors de l'étape E du EM, nous proposons une amélioration de l'algorithme du Forward-Backward qui diminue sa complexité temporelle. Nous démontrons que pour les MHMM-DF, l'étape E du EM, est réalisable avec une complexité seulement polynomiale. Nous avons aussi étendu l'algorithme de Viterbi à la structure de MHMM-DF. Celui-ci est utilisé pour retrouver la

suite d'états la plus probable dans une chaîne de Markov cachée. Cela permet de retrouver l'état de la banque de graines des adventices dans les champs. De la même manière que pour l'étape E du EM, la complexité de l'algorithme du Viterbi est linéaire en le nombre de patchs. Enfin, nous utilisons le modèle pour prédire la population non-dormante du pas de temps suivant.

Nous étudions la qualité de l'estimation des paramètres du modèle et celle de la prédiction ainsi que la restauration sur données simulées. Le modèle est appliqué aux données d'Epoisses qui recensent la flore levée adventice sur 90 champs pendant 17 ans. Les données d'Epoisses permettent d'étudier le comportement des adventices soumises à des méthodes de gestion différentes. Un travail du sol, un désherbage mécanique, un désherbage chimique ou un désherbage chimique et mécanique constituent les diverses méthodes de gestion utilisées sur le complexe d'Epoisses.

Les résultats obtenus sur les données d'Epoisses sont cohérents avec les données de la littérature. Ils mettent généralement en évidence la survie des graines comme le facteur prédominant dans la dynamique de la banque de graines des adventices. Les résultats selon la saison de la culture ont montré que la probabilité de germination est plus forte pendant les périodes favorables pour les adventices. De plus, la probabilité de colonisation provenant des patchs voisins est plus influente dans les saisons favorables à la grenaison des adventices. Plus de la moitié des prédictions de la flore levée par le MHMM-DF sont correctes, ce qui est bien meilleur qu'une prédiction au hasard correcte uniquement dans 20% des cas.

Chapitre 2

État de l'art

L'objectif de ce chapitre est de regrouper les connaissances en écologie, en modélisation et en statistiques nécessaires à la création d'un modèle qui représenterait la dynamique locale et régionale des adventices et serait facilement estimable. Pour cela, ce chapitre est composé de trois parties. La première partie regroupe les connaissances actuelles dans la littérature sur la dynamique des adventices. La deuxième partie s'intéresse aux articles modélisant la dynamique des espèces avec stade caché. Enfin, la troisième partie détaille les outils mathématiques nécessaires à la construction et à l'estimation d'un modèle avec une structure markovienne.

2.1 Que sait-on sur les adventices ?

Le cycle de vie des adventices se déroule en trois stades : graine, plantule ou plante adulte. Chaque stade remplit un rôle important dans la dynamique des adventices et une multitude de facteurs différents influence la survie d'un ou plusieurs stades de l'espèce. Le stade de plantule étant un stade intermédiaire entre la graine et la plante il ne sera pas utilisé lors de la modélisation.

2.1.1 Banque de graines

La banque de graines est souvent considérée comme une boîte noire en écologie car divers critères doivent être réunis pour qu'une graine puisse germer. Il faut que les conditions environnementales, comme la température, l'humidité, l'oxygène ainsi que la luminosité, lui soient favorables. Cependant, même quand elles sont adéquates, elles peuvent se révéler insuffisantes.

Un autre paramètre à prendre en compte est la dormance des graines. D'après Baskin and Baskin (2004), une graine est dite dormante si, en dépit de conditions climatiques et environnementales favorables sous une période de temps, elle n'a pas les capacités de germer. La dormance des graines permet de retarder la germination jusqu'à ce que les conditions environnementales soient idéales pour la croissance de la plante (Simpson, 1990) et celle-ci peut parfois durer plusieurs décennies. Notons que la dormance n'est pas engendrée par l'absence des conditions nécessaires pour la germination (Vleeshouwers et al., 1995). De plus, si les conditions de germination sont défavorables pendant trop longtemps mais que la graine survit, il est possible qu'une graine non dormante rentre dans un deuxième stade de dormance (Baskin and Baskin, 2004).

2.1.2 Les plantules et les plantes adultes

Pour leur croissance, les plantules ont besoin de lumière ainsi que de ressources qu'elles trouvent dans le sol. La croissance de la population de plantules est limitée par le surpeuplement, phénomène appelé densité-dépendance. En effet, pour les ressources, les plantules adventices sont en

compétition avec la culture semée et avec les autres adventices. Cependant, on peut négliger la compétition entre adventices dans un champ agricole étant donné la faible proportion d'adventices par rapport à la culture semée, si le champ est bien entretenu. Cet effet de densité-dépendance joue un rôle essentiel dans la dynamique des plantes (MacDonald and Watkinson, 1981). La survie des plantes adultes dépend des ressources utilisables dans le sol ainsi que de la lumière, du dioxyde de carbone, de l'eau et de la température.

2.1.3 La survie d'une population d'adventices

En plus des besoins spécifiques pour germer et grandir, les plantes sont souvent soumises au risque de prédation. Cette prédation peut avoir lieu sur les graines, par exemple avec les carabes, ainsi que sur les plantes. Cependant, les plantes sont connues pour avoir une stratégie de minimisation du risque d'extinction (Venable D (2007) ; Gremer and Venable (2014)) lié à des conditions environnementales inadéquates, à l'aide de la dormance des graines et la colonisation. Pour pouvoir survivre, une graine développe une stratégie selon deux axes : temporel par la dormance et spatial avec la colonisation (Cohen, 1966 ; Venable D and Brown, 1988). Si la dormance et la colonisation contribuent toutes deux à éviter des risques liés aux conditions environnementales, Renaud et al. (2013) ont prouvé qu'un compromis entre les deux existe grâce à un modèle théorique. Plus précisément, la dormance serait corrélée négativement à la colonisation. Cependant ce compromis n'a pas encore été validé empiriquement.

Au sein des champs agricoles, un facteur influençant particulièrement les chances de survie des adventices est le facteur humain, via les méthodes de gestion, rotation de culture ou type de désherbage.

2.1.3.1 Les types de dormance

D'après Baskin and Baskin (2004) les graines ont 5 types de dormance : la dormance morphologique, la dormance physiologique, la dormance morphophysiologique, la dormance physique et la combinaison de la dormance physique et physiologique.

La dormance morphologique représente la période du développement de l'embryon, suivi de la germination. La dormance physiologique empêche l'embryon de se développer et la graine de germer jusqu'à ce qu'un changement d'humidité, de température ou de luminosité, selon l'espèce, se produise. La dormance morphophysiologique correspond à la combinaison de la dormance morphologique et physiologique. La dormance physique est causée par la résistance d'une ou plusieurs enveloppes imperméables de la graine. Elle est arrêtée une fois que les enveloppes se sont dégradées. La dormance physique et physiologique correspond à la combinaison de la dormance physique et physiologique. Ces différents types de dormance font varier le temps de dormance selon l'espèce. Ainsi, chaque type de dormance pourrait avoir un effet différent sur la survie d'une graine.

2.1.3.2 Différentes types de colonisation

Une fois la reproduction réalisée, les graines sont produites et dispersées. Cette dispersion a un effet sur la survie de l'espèce. Si aucune dispersion n'est faite la survie des graines ne dépendra que des conditions environnementales et climatiques locales. Ainsi, si des conditions climatiques et environnementales adéquates ne se présentent pas pendant plusieurs années, l'espèce risque de s'éteindre. Cependant si la dispersion des graines a lieu, leur survie dépendra des variations environnementales. Les facteurs environnementaux et climatiques qui influencent la colonisation sont le vent, les cours d'eau, la faune et l'activité humaine. De plus, la distance de dispersion dépend des caractéristiques de l'espèce. En effet, la taille, le poids et la forme des graines ainsi que la hauteur de la plante et les périodes de reproduction peuvent influencer la dissémination.

2.1.4 Les méthodes de gestion des cultures

Les adventices ne sont pas la seule source de nuisance dans un champ agricole. Certaines espèces d'animaux et certaines maladies peuvent ravager les champs. Au cours des années, l'homme a trouvé de nombreuses méthodes pour remédier aux nuisances dans les champs agricoles. Le taux de réussite de chaque méthode de gestion peut varier selon le type de bio-agresseur, l'environnement et le climat, la saison et la rigueur avec laquelle la méthode est appliquée. Le choix de la méthode de gestion est souvent laissé à l'agriculteur. La méthode de gestion utilisée aura des répercussions sur le plan économique, écologique et toxicologique.

Les méthodes de gestion des cultures peuvent être scindées en deux grands groupes : celles qui présentent des actions directes sur la population de bio-agresseurs et celles qui présentent des actions indirectes sur la population de bio-agresseurs. Les méthodes à actions indirectes ont pour objectif premier de placer les plantes cultivées dans les meilleures conditions possibles alors que les méthodes à actions directes sont des méthodes de lutte contre les bio-agresseurs. Les méthodes de lutte directe contre les adventices se font soit par action directe sur le stock semencier soit par action sur le développement des adventices et la grenaison.

2.1.4.1 Les méthodes à actions indirectes sur les adventices

Les actions indirectes incluent toutes les méthodes d'amélioration du sol. Le sol peut être amélioré par des méthodes de drainage, en utilisant un amendement tel que le fumier. Cela comprend aussi l'écartement du semis dans le champ. Les méthodes d'actions indirectes sur les bio-agresseurs ne se limitent pas au travail du sol. Cela comprend aussi la sélection de l'espèce cultivée, l'irrigation contrôlée, l'entretien des abords de la parcelle, la rotation de la culture et la date de semis. L'espèce cultivée peut être choisie en fonction des résistances de l'espèce face à certains types de bio-agresseurs. La micro-irrigation est une méthode d'irrigation contrôlée, qui consiste à nourrir la plante cultivée par goutte-à-goutte. Cette méthode est utilisée dans les zones arides afin de réduire l'utilisation d'eau et d'engrais. La rotation de la culture permet d'éviter une surcroissance d'une certaine espèce de bio-agresseur.

2.1.4.2 Les méthodes à actions directes sur les adventices

Les méthodes à actions directes sur les bio-agresseurs ont pour objectif de réduire leur influence sur la culture semée. Elles peuvent être regroupées en trois grandes classes : les méthodes chimiques, physiques et biologiques.

2.1.4.3 Méthodes de lutte chimique

Les méthodes chimiques correspondent à l'utilisation d'une substance chimique afin de lutter contre les bio-agresseurs. Ce type de méthode peut être utilisé pour exercer une action sur l'un des processus vitaux des bio-agresseurs, pour freiner la croissance de ces bio-agresseurs, pour tuer les bio-agresseurs ou pour les faire fuir. Ces méthodes sont utilisées pour protéger les plantes cultivées des bio-agresseurs. Cependant, le recours aux méthodes chimiques a un impact sur des organismes non-cibles ainsi que sur l'écosystème. De plus, l'utilisation répétée de produits chimiques diminue leur efficacité au fil du temps car les bio-agresseurs deviennent de plus en plus résistants à ces produits. De ce fait, l'utilisation de pesticides devrait être remplacée sur le long terme par d'autres méthodes de gestion. Ces méthodes peuvent freiner la croissance de l'adventice, stériliser les adventices voire les tuer.

2.1.4.4 Méthodes de lutte physique

Beaucoup de méthodes physiques existent afin de lutter contre les bio-agresseurs. Les méthodes physiques de protection consistent à protéger la plante cultivée à l'aide de filets ou de bâche afin

de bloquer tout bio-agresseur susceptible d'y accéder. La bâche opaque recouvrant le sol empêche la lumière de passer et entrave la survie des adventices après la germination. Elle sert aussi à désinfecter le sol. D'une part, injecter de la vapeur sous la bâche peut mettre certains bio-agresseurs hors d'état de nuire, d'autre part, elle permet une augmentation de la température. Cette technique appelée solarisation, n'est pas la seule technique de lutte thermique contre les bio-agresseurs. Il existe aussi l'élimination des adventices par feux contrôlés Quintana-Ascencio et al. (2011). Pour lutter contre les adventices, il est aussi possible d'utiliser un désherbage mécanique à l'aide d'outils comme la herse étrille, la houe rotative et les bineuses. La récupération des menues pailles à l'arrière de la moissonneuse-batteuse contribue à réduire les populations de graines dans le champ. Cette méthode permet de récupérer les graines pendant la récolte. Une fois récoltées, les graines peuvent être brûlées ou broyées. Cette méthode détruit environ 95% des graines d'ivraie Stokstad (2013). On recourt aussi à des techniques de faux semis, qui consistent à préparer un lit de semences afin d'inciter les graines à germer. Une fois les graines d'adventices germées une méthode de désherbage est appliquée. Le semis de la culture s'opère juste après ce désherbage.

2.1.4.5 Méthodes de lutte biologique

Les méthodes biologiques englobent l'ensemble des méthodes de protection des végétaux qui utilisent les mécanismes et interactions régissant les relations entre espèces dans le milieu naturel. En d'autres termes, le principe est fondé sur la gestion des équilibres des populations d'agresseurs plutôt que sur leur éradication. La gestion par méthodes biologiques est pratiquée par le biais de micro ou macro-organismes. Les macro-organismes regroupent les espèces comme les invertébrés, les insectes, les acariens ou même les nématodes. L'introduction de ces espèces dans un champ d'adventices permet d'insérer un prédateur ou un parasite naturel du bio-agresseur ciblé. Les micro-organismes sont des champignons, des bactéries et des virus utilisés pour protéger les cultures contre les ravageurs et les maladies. Ces méthodes peuvent avoir pour cibles les graines des adventices ou directement les plantes adventices.

2.1.5 Discussion

Afin de modéliser la dynamique locale et régionale des adventices, il faut déterminer les processus qui gouvernent la dynamique et les facteurs ayant un impact sur ces processus. Les processus indispensables à la persistance des adventices dans l'environnement ne peuvent pas être négligés lors de la modélisation. Pour les adventices, il est essentiel de modéliser les processus de dormance, de non-dispersion de graines, de colonisation ainsi que le processus de germination. Cependant, certains de ces processus dépendent de facteurs environnementaux et biologiques comme le vent et les animaux. Afin d'inclure ces facteurs dans le modèle, il est possible d'en étudier à l'aide de données dont nous disposons. Cependant, quand les facteurs ne sont pas liés à un facteur humain, les données peuvent être difficiles à obtenir. De plus, certains de ces facteurs peuvent être négligeables par rapport à d'autres. En revanche, les données concernant les facteurs humains, tels que la culture semée ou le type de désherbage, sont faciles à acquérir. Ainsi, pour l'application du cadre MHMM-DF aux données d'Epoisses, nous montrerons comment prendre en compte l'influence de la culture et de la pratique agricole sur la survie de la banque de graines ainsi que la colonisation des adventices.

Avant de plonger dans la modélisation, la prochaine section expose les différents modèles déjà existants pour modéliser la dynamique d'espèces avec stade caché.

2.2 La modélisation de la dynamique des espèces avec stade caché

Dans cette section nous allons détailler les modèles de la littérature qui prennent en compte la dynamique des espèces avec stade caché. Cette section est organisée de façon à avoir une colonisation qui devient de plus en plus réaliste au cours de la bibliographie. Chaque modèle est détaillé selon ses limites en termes de prise en compte de la colonisation, de facilité d'estimation, du nombre d'états pour les variables et aussi des éventuelles hypothèses restrictives imposées par le modèle. Toutes les notations utilisées sont identiques à celles décrites dans les articles associés. Les modèles présentés ont été créés dans des buts divers : trouver le nombre d'individus ou l'équilibre du modèle, estimer les paramètres associés à la dynamique de l'espèce, chercher la stratégie optimale qui maximise les chances de survie de l'espèce, prédire la distribution de la population au temps suivant ou encore déterminer la dynamique la plus appropriée pour une espèce à l'aide d'estimation.

Ayant pour objectif d'estimer la dynamique locale et régionale d'une espèce avec stade caché à l'aide de données partielles, nous avons favorisé les modèles utilisés pour de l'estimation qui considéraient la banque de graines non observable. En revanche, on ne présente pas ici toute la bibliographie existante sur les modèles n'ayant pas pour but l'estimation. Seuls ceux présentant une originalité dans leur approche ont été sélectionnés. La plupart d'entre eux utilisent des équations différentielles pour modéliser la dynamique d'une espèce avec dormance.

2.2.1 Modèle de métapopulation de Levins

Le modèle de Levins est fréquemment utilisé pour étudier la dynamique d'une espèce à travers plusieurs patchs. Il n'utilise que 2 paramètres, un paramètre de colonisation et un paramètre d'extinction. Soit N la fraction des patchs occupés et c un taux constant de propagules générées par les N patchs occupés pendant un temps dt. Le nombre d'habitats occupés par l'espèce au sein de la métapopulation N au cours du temps est donné par la résolution de l'équation différentielle :

$$\frac{dN}{dT} = cN(1-N) - eN$$

où e est la probabilité que l'espèce meure au sein d'un patch. L'avantage du modèle de Levins réside dans sa simplicité et dans la facilité à pouvoir l'utiliser sur différents types d'espèces.

Limites

Le modèle de Levins (Levins et al., 1969) ne convient pas aux espèces ayant un stade de dormance dans leur cycle de vie. Si néanmoins on choisit d'appliquer le modèle de Levins sur la population non dormante et que celle-ci s'éteint, le modèle va supposer que l'espèce n'est plus présente dans le patch alors qu'elle peut toujours être présente sous forme dormante. Ainsi une fois que la population non dormante réapparaît dans le patch, le modèle de Levins va supposer que cela est dû à la colonisation. Fréville et al. (2013) ont montré que le modèle de métapopulation sur les espèces avec dormance surestimait les paramètres de colonisation et d'extinction.

2.2.2 Modélisation de la dynamique locale sans colonisation

Cette section détaille les modèles qui ne prennent pas en compte la colonisation dans la dynamique des espèces avec stade caché, c'est-à-dire qu'aucune spatialisation n'est prise en compte.

2.2.2.1 Modèle de Han et al

Han et al. (2014) modélisent la dynamique des plantes avec dormance à partir du nombre de graines dans la banque de graines. C'est un modèle déterministe qui permet de représenter plusieurs

niveaux de densité-dépendance dans la dynamique de l'espèce. La prise en compte de la densité-dépendance permet une étude détaillée de la dynamique locale qui considère la compétition entre individus au sein de la même espèce. Voici les paramètres :

— $H(t)$ est le rendement moyen du nombre de graines produites par chaque plante pendant l'année t,
— $S(t)$ est le nombre de graines dans la banque de graines pendant l'année t,
— g est le taux de germination,
— d est le taux de mortalité des graines dans la banque de graines,
— Q est la capacité de charge,
— $1/k$ est le risque d'avoir une germination retardée,
— b est un facteur de proportionnalité.

$$S(t+1) = S(t)H(t)g\left(\frac{(1+bgS(t)}{Qk}\right)^{-k} + S(t)(1-g)(1-d)$$

où $\left(\frac{(1+bgS(t)}{Qk}\right)^{-k}$ est la fonction de densité-dépendance. Avec leur modèle, les auteurs cherchent la stratégie évolutivement stable de l'espèce, c'est-à-dire le nombre de graines tel qu'au temps suivant, le nombre de graines reste inchangé.

Limites

Le modèle est déterministe. Il n'inclut pas de colonisation et ne permet pas de faire d'estimation sans information sur la banque de graines.

2.2.2.2 Modèle de Quintana-Ascencio et al

Quintana-Ascencio et al. (2011) modélisent, de façon déterministe, la dynamique des plantes sur des champs qui sont annuellement entièrement brûlés. Ainsi, leur dynamique peut être vue comme une dynamique de plantes annuelles. Quintana-Ascencio et al. (2011) utilisent ce modèle pour étudier l'espèce *Warea carteri* et comparent les trajectoires simulées avec les trajectoires observées. Le modèle considère 4 stades de vie : les graines de moins d'un an, les graines de plus d'un an, les plantules et les plantes adultes. Dans le modèle, il n'y pas d'interaction d'une année sur l'autre de flore levée à flore levée. Ainsi il y a forcément dormance des graines. Les paramètres du modèle sont :

— f la fécondité,
— s le taux de survie des graines,
— d_1, d_2, d_3 les taux de dormance des graines selon leur âge,
— g_1, g_2, g_3 les taux de germination selon l'âge des graines,
— z_1, z_2 les taux de survie des plantules avant et après leur détection.

Les dépendances sont représentées dans la figure 2.1. Voici comment est représentée la transition d'états :

$$X(t+1) = X(t) \times A$$

où $X(t+1)$ représente l'état de chaque stade et A est la matrice de transition définie par

$$A = \begin{array}{c} \\ \mathrm{BG}_0 \\ \mathrm{BG}_1 \\ \text{Plantule} \\ \text{plant adult} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{BG}_0 & \mathrm{BG}_1 & \text{Plantule} & \text{plant adult} \end{array} \\ \begin{pmatrix} 0 & SBK_1 & SDS_1 & 0 \\ 0 & SBK_2 & SDS_2 & 0 \\ 0 & 0 & 0 & SUR \\ FSB & 0 & FSD & 0 \end{pmatrix} \end{array}$$

Quintana-Ascencio et al ont un modèle avec dormance obligatoire. Cependant ils ont étendu leur modèle à un modèle qui peut ne pas avoir de dormance. On note D la proportion de graines

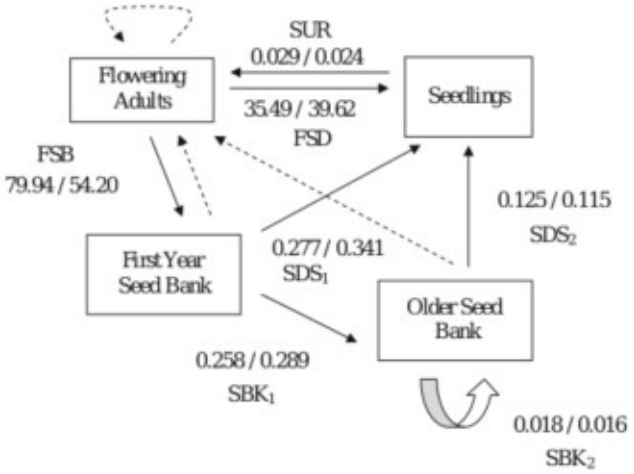


FIGURE 2.1 – Figure issue de Quintana-Ascencio et al. (2011). $FSB = fs^2d_1, SBK_1 = s^2d_2, SBK_2 = sd_3, FSD = fsz_1g_1, SDS_1 = sz_1g_2, SDS_2 = z_1g_3, SUR = z_1z_2$. Soit BG_0 et BG_1 le nombre de graines dans la banque de graines âgées respectivement de moins d'un an et de plus d'un an.

qui germent sans dormance. Ainsi, la matrice de transition associée est :

$$\begin{array}{c} \\ BG_0 \\ BG_1 \\ \text{Plantule*} \\ \text{plant adulte} \end{array} \begin{array}{c} \begin{array}{cccc} BG_0 & BG_1 & \text{Plantule} & \text{plant adult} \end{array} \\ \begin{pmatrix} 0 & SBK_1 & SDS_1(1-D) & SDS_1 \times D \times SUR \\ 0 & SBK_2 & SDS_2(1-D) & SDS_2 \times D \times SUR \\ 0 & 0 & 0 & SUR \\ FSB & 0 & FSD(1-D) & FSD \times D \times SUR \end{pmatrix} \end{array}$$

Leurs résultats montrent que le retardement dans la germination est peut-être la cause du cycle biannuel des populations. Leur modèle est utilisé afin de simuler les trajectoires des plantes et de les comparer aux trajectoires réelles.

Limites
Le modèle ne présente pas de colonisation.

2.2.2.3 Modèle de Jarry et al

Jarry et al. (1995) modélisent la dynamique des plantes annuelles à partir du nombre d'individus dans les populations dormantes et non-dormantes. Le modèle est déterministe et ne présente que 4 paramètres. Cependant, le modèle repose sur 2 hypothèses :

1. Le paramètre démographique des graines est constant et indépendant de l'âge des graines ;
2. Les graines dormantes dans la banque de graines ne sont pas soumises au risque de mortalité.

L'évolution des populations est calculée de la manière suivante :

$$X(t+1) = AX(t)$$

où $X(t) = (x_1(t), x_2(t))$ représente l'état des populations dormantes et non dormantes avec $x_1(t), x_2(t)$ le nombre de graines dans la banque de graines et le nombre de plantes adultes au

temps t. La matrice de transition A entre deux pas de temps est la suivante :

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

où $a_{11}, a_{12}, a_{21}, a_{22}$ représentent respectivement le taux de graines qui restent dormantes, le taux de graines produites par plante et entrant dans la banque de graines, le taux de graines qui germent et survivent jusqu'à l'âge adulte et le taux de graines produites par adulte qui germent directement dans l'année qui suit. Le première ligne de la matrice A décrit les processus qui font varier le nombre de graines dans la banque de graines et la deuxième ligne décrit les processus qui font varier le nombre de plantes adultes (voir 2.2).

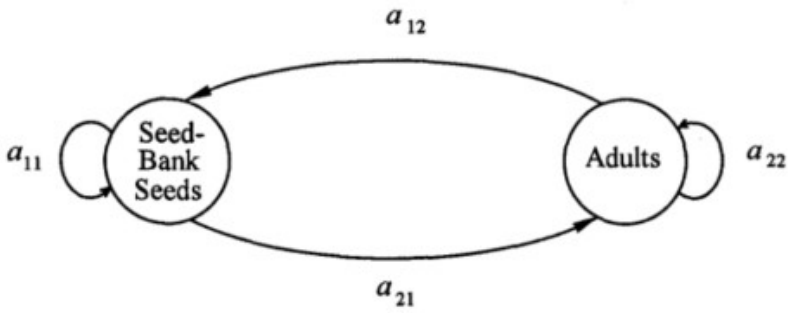


FIGURE 2.2 – Figure issue de Jarry et al. (1995)

Les coefficients de la matrice A s'expriment en fonction des différents processus biologiques :

$$\begin{array}{rcl} a_{11} & = & a \\ a_{12} & = & af \\ a_{21} & = & (1-a)gs \\ a_{22} & = & (1-a)gsf \end{array}$$

où a correspond au taux de graines restant dormantes, f au taux de graines produites, g au taux de germination et s au taux de survie d'une plante adulte.

Jarry et al ont aussi montré que le modèle peut être étendu pour tenir compte de l'effet de la densité. Cette extension fait dépendre f et s de la densité-dépendance de l'espèce au temps t. Jarry et al ont utilisé leur modèle pour de l'estimation sur données réelles pour l'espèce Sesbania vesicaria.

Limites

Les hypothèses d'âge et de mortalité sur les graines sont très restrictives. De plus, le modèle est déterministe et ne considère pas de colonisation. Les méthodes d'estimation supposent que la banque de graines est observable.

2.2.2.4 Modèle de Borgy et al

Borgy et al. (2015) s'intéressent dans un premier temps au nombre exact d'individus pour ensuite construire un modèle en classes d'individus, dans le but d'utiliser des données d'abondance sur la flore levée. Il s'agit d'un modèle de Markov caché où les plantes adultes sont observables et la banque de graines est cachée. L'intérêt du modèle est qu'il permet d'étudier la dynamique des adventices en considérant les méthodes de gestion employées dans le champ.

Décrivons d'abord le modèle au niveau du nombre d'individus. Chaque graine dans la banque de graines a une probabilité σ de germer et de survivre jusqu'à l'âge adulte. Ainsi le nombre de

plantes X^{t+1} à l'année $t+1$ est le résultat de la germination de graines et suit une loi Binomiale de paramètres (Y^t, σ), où Y^t est le nombre de graines dans la banque de graines. On a

$$Y^{t+1} = Y^t - X^{t+1} - Y_m^{t+1} + Y_p^{t+1}$$

où Y_m^{t+1} est le nombre de graines mortes et Y_p^{t+1} est le nombre de graines produites par X^{t+1}. Y_m^{t+1} suit une loi Binomiale de paramètres $(Y^t - X^{t+1}, 1-s)$ où s est un paramètre de survie. Le modèle suppose que Y_p^{t+1} suit une loi de Poisson de paramètre $X^{t+1}\phi$.

Les variables Y^t et X^{t+1} sont ensuite rendues discrètes en classes d'abondance. On note $I_{C_{x^t}}$ l'intervalle des valeurs possibles du nombre de plantes levées dans la classe C_{x^t} et $I_{C_{y^t}}$ l'intervalle des valeurs possibles du nombre de graines dans la classe C_{y^t}. Alors on peut exprimer les probabilités de transition du modèle à classes d'abondance de la manière suivante :

$$\mathbb{P}(C_{x^{t+1}}|C_{y^t}) = \sum_{x^{t+1} \in I_{C_{x^{t+1}}}} \sum_{y^t \in I_{C_{y^t}}} \mathbb{P}(y^t|C_{y^t})\mathbb{P}(x^{t+1}|y^t)$$

$$\mathbb{P}(C_{y^{t+1}}|C_{x^{t+1}}, C_{y^t}) = \sum_{x^{t+1} \in I_{C_{x^{t+1}}}} \sum_{y^t \in I_{C_{y^t}}} \sum_{y^{t+1} \in I_{C_{y^{t+1}}}} \sum_{y_m^{t+1} \in I_{C_{y_m^{t+1}}}} \mathbb{P}(y_m^{t+1}|y^t, x^{t+1})\mathbb{P}(y^t|C_{y^t})...\mathbb{P}(x^{t+1}|C_{x^{t+1}})\mathbb{P}(y_p^{t+1}|x^{t+1})$$

Le mode de gestion est déterminé par la culture semée. Le modèle considère 3 paramètres par méthode de gestion (σ_a, s_a, ϕ_a) où l'indice a indique la culture. Le graphe des dépendances de ce modèle de Markov caché est représenté sur la figure 2.3. On remarque dans la figure 2.3 que la flèche de $C_{x^{t+1}}$ vers $C_{y^{t+1}}$ n'est pas présente dans une simple chaîne de Markov cachée et elle représente l'apport de nouvelles graines dans la banque de graines. Ainsi on appellera cette structure une chaîne de Markov cachée avec retour de données.

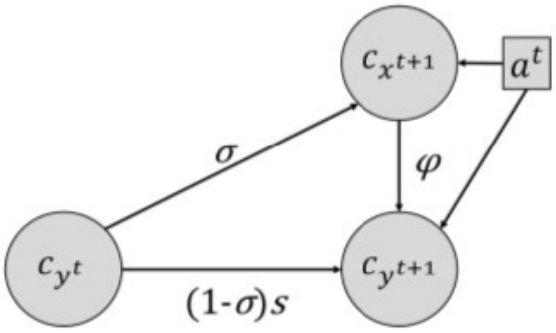


FIGURE 2.3 – Figure issue de Borgy et al. (2015)

Borgy et al ont utilisé leur modèle pour estimer la dynamique de plusieurs types d'espèces en fonction de la culture sur données réelles.

Limites

Le modèle n'inclut pas de colonisation qu'elle soit spatiale ou par pluie de graines. Le fait de modéliser les populations à partir des individus pour ensuite les agréger en classes d'abondance rallonge le temps de calcul des probabilités de transition car il repose sur des simulations de Monte Carlo.

2.2.3 Modélisation de la dynamique locale avec colonisation

L'un des problèmes de modélisation de la dynamique locale sans colonisation est qu'une fois que les populations dormantes et non dormantes meurent, l'état de l'espèce dans le patch est

irréversible. En d'autres termes, l'état d'extinction d'une espèce est un état absorbant si l'on ne considère pas de colonisation. Dans cette section la colonisation est présente sous forme de pluie de graines. A chaque pas de temps, la probabilité que le patch soit colonisé par une source voisine non localisée géographiquement est soit constante, soit considérée comme une variable aléatoire qui sera indépendante des patchs voisins. De plus, chaque patch est soumis à la même colonisation.

2.2.3.1 Modèle de Regan et al

Le modèle de Regan et al. (2006) est simple et n'utilise que des données d'observation de type présence/absence sur la flore levée éventuellement bruitées. Ainsi, la banque de graines n'est pas observée. Le modèle suppose que de nouvelles graines sont forcément créées s'il y a présence de flore levée. Le modèle prend en compte la probabilité de détection de la flore levée. S est une chaîne de Markov et correspond à la population dans le patch. Elle peut valoir S_e dans un patch vide, S_s dans un patch où seules les graines sont présentes et S_w dans un patch où les plantes et les graines adventices sont présentes. On note Z la variable d'observation du modèle. Elle vaut z_a si aucune population n'est observée et z_p si une population est observée dans le patch. Les transitions autorisées entre les états sont schématisées sur la figure 2.4.

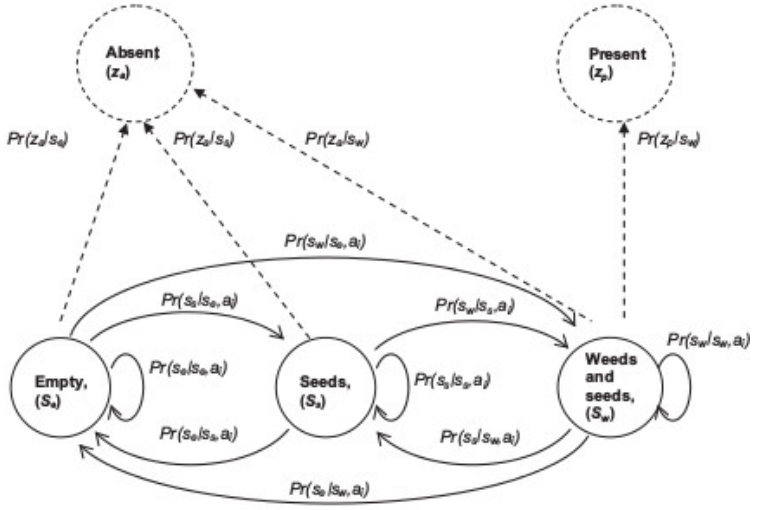


FIGURE 2.4 – Figure issue de Regan et al. (2006)

Les probabilités de transition entre chaque état du patch peuvent être décrites grâce à 4 paramètres. On note g la probabilité de germination et c la probabilité de colonisation. La probabilité ρ_1 (respectivement ρ_2) correspond à l'extinction des graines dans la banque de graines sans (respectivement avec) création de nouvelles graines. Le paramètre a définit la culture.

$$\begin{array}{rcc}
\mathbb{P}(S_e|S_e,a) & = & 1-c \\
\mathbb{P}(S_s|S_e,a) & = & c(1-g) \\
\mathbb{P}(S_w|S_e,a) & = & cg \\
\mathbb{P}(S_e|S_s,a) & = & \rho_1 \\
\mathbb{P}(S_s|S_s,a) & = & (1-\rho_1)(1-g) \\
\mathbb{P}(S_w|S_s,a) & = & (1-\rho_1)g \\
\mathbb{P}(S_e|S_w,a) & = & \rho_2 \\
\mathbb{P}(S_s|S_w,a) & = & (1-\rho_2)(1-g) \\
\mathbb{P}(S_w|S_w,a) & = & (1-\rho_2)g
\end{array}$$

Ce modèle est un POMDP (Partially Observed Markovian Descision Process) (Kaelbling et al., 1998). Un POMDP est un modèle pour la conception par optimisation de stratégie de gestion optimale en contexte temporel. Dans cet article, les différentes stratégies résident dans le choix de la culture semée.

Limites

Le modèle suppose que des graines sont forcément produites dès que des plantes adultes sont présentes. De plus, le modèle ne prend pas en compte la colonisation entre patchs voisins, c'est-à-dire que la population non-dormante d'un patch n'influence pas la population dormante d'un autre patch.

2.2.3.2 Modèle de Pluntz et al

Le modèle de Pluntz et al. (2018) est un HMM avec retour de données qui a pour but d'estimer les paramètres de la dynamique des adventices. Il utilise des données réelles de type présence/absence uniquement sur la flore levée pour faire de l'estimation à l'aide de l'algorithme EM. Le graphe de dépendance du modèle est montré dans la figure 2.5 :

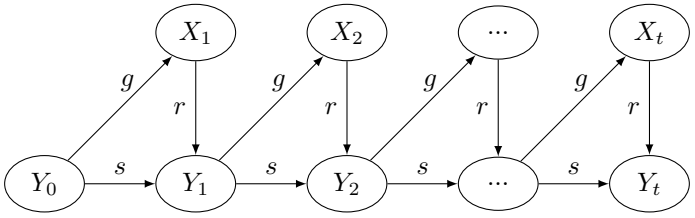


FIGURE 2.5 – Le modèle de Pluntz et al

La variable de Bernoulli $X_t \in \{0, 1\}$ vaut 1 si l'on observe de la flore levée au temps t et 0 sinon. La variable de Bernoulli $Y_t \in \{0, 1\}$ vaut 0 si la banque de graines est vide au temps t et 1 sinon.

On note g la probabilité de germination, r la probabilité de reproduction, s la probabilité de survie de la banque de graines et c la probabilité de colonisation par pluie de graines. Voici comment les probabilités de transition sont modélisées :

$$\mathbb{P}(X_{t+1} = 1 | Y_t = y_t) = g y_t$$

$$\mathbb{P}(Y_{t+1} = 1 | Y_t = y_t, X_{t+1} = x_{t+1}) = 1 - (1 - s y_{t-1})(1 - r x_{t+1})(1 - c)$$

L'estimation est rapide avec l'algorithme EM. Le modèle a été validé sur données réelles et permet de déterminer les rôles respectifs de la banque de graines et de la colonisation dans la dynamique des espèces adventices. Nous avons montré que les adventices ont une compatibilité avec une ou plusieurs cultures. Ainsi, des adventices de printemps ont beaucoup plus de mal à survivre dans un champ avec une culture d'hiver qu'une culture d'été.

Limites

Les populations d'un patch ne dépendent que de la population du patch au temps précédent. Par conséquent, le modèle ne considère pas d'interaction entre patchs. De plus, le modèle n'est pas identifiable avec 4 paramètres et nécessite d'avoir un paramètre fixé. En fixant le paramètre de reproduction r à 1, le modèle impose que si la flore levée est présente dans un patch alors la banque de graines au temps suivant est forcément présente.

2.2.3.3 Modèle de David et al

David et al. (2010) proposent un modèle stochastique utilisé pour estimer la dynamique des plantes annuelles à partir du nombre d'individus. Le modèle est assez complet puisqu'il intègre

plusieurs processus biologiques comme la reproduction, l'immigration de graines, la survie des graines dans la banque de graines ainsi que la croissance des plantes. Tous ces processus sont modélisés selon des lois connues et deux types de modélisation sont utilisés. Détaillons une des modélisations proposées par David et al. L'autre utilise une loi Binomiale négative afin de modéliser la production de graines.

On distingue 3 groupes de population au temps t : les graines dans la banque de graines $S_{i,t}$ d'âge i, les rosettes $R_{i,t}$ provenant de graines d'âge i et les plantes adultes F_t. Le nombre de graines produites par l'ensemble des plantes suit une loi de Poisson de paramètre M_t où M_t suit une loi gamma de paramètres $(\alpha_m F_{t-1}, \theta)$. On note $m = \alpha_m/\theta$ l'espérance du nombre de graines produites par plante. Le nombre de graines rentrant dans la banque de graines via colonisation suit une loi de Poisson de paramètre U_t où U_t qui suit une loi gamma de paramètres (α_u, θ). On note $u = \alpha_u/\theta$ l'espérance du nombre de graines rentrant via colonisation. Le nombre de plantes matures F_t suit une loi Binomiale de paramètres $(\sum_i R_{i,t}, c)$. On note a_i la probabilité qu'une graine d'âge i survive 1 an. On note e_i la probabilité qu'une graine d'âge i germe.

La variable aléatoire $S_{0,t+1}$ est égale à la somme du nombre de graines issues de la colonisation et du nombre de graines produites par les plantes locales. La variable $S_{1,t+1}$ est égale au nombre de vieilles graines qui ne germent pas et survivent. Le nombre de plantules $R_{i,t}$ est égal au nombre de graines d'âge i qui germent. Le nombre de vieilles graines d'âge i qui meurent est noté $D_{i,t}$. Ainsi $(S_{i,t+1}, R_{i,t}, D_{i,t}), D_{i,t}|S_{i,t}$ suit une loi multinomiale de paramètres $(S_{i,t}, a_i, e_i, 1-a_i-e_i)$ où $S_{i,t}$ est le nombre d'états de la multinomiale. Voici le graphe associé au modèle de David et al :

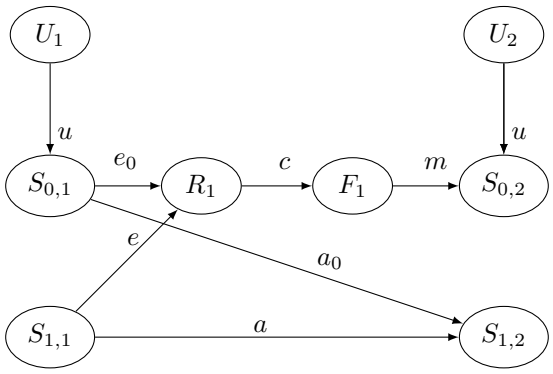


FIGURE 2.6 – Le modèle de David et al

Ce modèle a été utilisé pour estimer la dynamique du colza dans des champs d'agriculture.

Limites

Le modèle détaille beaucoup de processus. Cependant il n'est pas identifiable en considérant tous les paramètres inconnus. Il faut en fixer au moins 2 pour estimer correctement tous les paramètres. De plus, l'algorithme d'estimation utilisé parvient à bien estimer les paramètres quand le modèle ne présente pas de dormance. Néanmoins, la convergence de l'algorithme n'est pas garantie pour un modèle avec dormance. Le modèle ne présente pas de colonisation inter-parcelles.

2.2.3.4 Modèle de Lamy et al

Le modèle de Lamy et al. (2013) décrit la dynamique des espèces avec état cryptique. Une espèce est dite dans un état cryptique si elle est difficilement observable. Dans cet article, les auteurs considèrent le cas particulier où l'état cryptique dépend des conditions climatiques. Leur modèle a été utilisé sur des données de type présence/absence d'escargots d'eau dans des zones arides.

Quand les conditions climatiques sont adéquates (W), les escargots peuvent vivre normalement. Cependant, quand les conditions climatiques ne sont pas adéquates (D), l'espèce est contrainte de vivre sous forme dormante dans le sol en attendant la future période adéquate. Pendant ces périodes inadéquates, les escargots sont difficilement observables. Le modèle utilise 5 paramètres :

— ϕ_W la probabilité de survie dans des conditions adéquates,
— ϕ_D la probabilité de survie dans des conditions inadéquates,
— ρ_W la probabilité de colonisation dans des conditions adéquates,
— ρ_D la probabilité de colonisation dans des conditions inadéquates,
— p_W la probabilité de détection dans des conditions adéquates.

Le modèle est un HMM avec une matrice de transition dépendant de la transition entre les saisons avec des données de type présence/absence sur la population non-dormante de l'espèce. On définit $q_{i,t}$ la variable qui représente les conditions climatiques et qui vaut 1 quand elles sont inadéquates dans le champ i au temps t. Ainsi 4 matrices de transition sont définies par $P_{DD}, P_{DW}, P_{WD}, P_{WW}$ de la manière suivante :

$$P_{WW} = \begin{array}{c} \\ \text{Absence} \\ \text{Présence} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absence} & \text{Présence} \end{array} \\ \begin{pmatrix} 1-\rho_W & \rho_W \\ (1-\phi_W)(1-\rho_W) & \phi_W + (1-\phi_W)\rho_W \end{pmatrix} \end{array}$$

$$P_{WD} = \begin{array}{c} \\ \text{Absence} \\ \text{Présence} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absence} & \text{Présence} \end{array} \\ \begin{pmatrix} 1-\rho_D & \rho_D \\ (1-\phi_W)(1-\rho_D) & \phi_W + (1-\phi_W)\rho_D \end{pmatrix} \end{array}$$

$$P_{DW} = \begin{array}{c} \\ \text{Absence} \\ \text{Présence} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absence} & \text{Présence} \end{array} \\ \begin{pmatrix} 1-\rho_W & \rho_W \\ (1-\phi_D)(1-\rho_W) & \phi_D + (1-\phi_D)\rho_W \end{pmatrix} \end{array}$$

$$P_{DD} = \begin{array}{c} \\ \text{Absence} \\ \text{Présence} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absence} & \text{Présence} \end{array} \\ \begin{pmatrix} 1-\rho_D & \rho_D \\ (1-\phi_D)(1-\rho_D) & \phi_D + (1-\phi_D)\rho_D \end{pmatrix} \end{array}$$

avec

$$P = q_{i,t}q_{i,t+1}P_{DD} + q_{i,t}(1-q_{i,t+1})P_{DW} + (1-q_{i,t})q_{i,t+1}P_{WD} + (1-q_{i,t})(1-q_{i,t+1})P_{WW}.$$

La détection de l'espèce n'est pas systématique et dépend de la saison. La matrice d'émission du HMM dans une saison inadéquate est :

$$P_{OD} = \begin{array}{c} \\ \text{Absence} \\ \text{Présence} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absence} & \text{Présence} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \end{array}$$

La probabilité d'émission dans une saison adéquate est :

$$P_{OW} = \begin{array}{c} \\ \text{Absence} \\ \text{Présence} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absence} & \text{Présence} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1-p_W & p_W \end{pmatrix} \end{array}$$

Le modèle est utilisé pour faire de l'estimation de la dynamique des escargots d'eau en zones arides à l'aide de méthode de MCMC.

Limites

L'espèce est modélisée par des données binaires de type présence/absence. Le modèle suppose que l'état cryptique est lié à la saison, ce qui n'est pas le cas pour la dormance des graines. De plus, il nécessite de savoir à chaque pas de temps si les conditions environnementales sont adéquates ou non. Enfin la colonisation ne prend pas en compte la présence de l'espèce dans les patchs voisins.

2.2.3.5 Modèle de Fréville et al

Le modèle de Fréville et al. (2013) cherche à décrire la dynamique des plantes annuelles. Les populations sont modélisées par présence/absence. Les populations dormantes sont non observables et les populations non dormantes sont observables. Le modèle a 4 paramètres :

— c la probabilité de colonisation par pluie de graines,
— g_0 la probabilité de germination des graines entre le temps t et le temps $t+1$,
— g_1 la probabilité de dormance et de germination du temps t au temps $t+2$,
— d la probabilité de perturbation.

La probabilité de perturbation d est une probabilité d'extinction des populations non dormantes indépendante de la dynamique de l'espèce. Un patch peut se voir attribuer un des 4 états suivants :

— A_A = Absence de population dormante et non-dormante,
— P_P = Présence de population dormante et non-dormante,
— A_P = Absence de population non-dormante et présence de population dormante,
— P_A = Présence de population non-dormante et absence de population dormante.

Leur modèle suppose que si une population non dormante est présente au temps t alors des graines seront présentes dans la banque de graines au temps $t+1$. On suppose aussi que les plantes produisent des graines qui germent soit l'année prochaine soit l'année d'après. Ces deux hypothèses impliquent que l'état de la banque de graines est toujours connu. Ainsi leur modèle peut être étudié à l'aide de chaînes de Markov. La matrice de transition d'un patch est définie de la sorte :

$$\begin{array}{c} \\ A_A \\ A_P \\ P_A \\ P_P \end{array} \begin{array}{c} \begin{array}{cccc} A_A & A_P & P_A & P_P \end{array} \\ \begin{pmatrix} 1-c(1-d) & 0 & c(1-d) & 0 \\ (1-c)(1-g_1)+d(c+g_1-cg_1) & 0 & (1-d)(c+g_1-cg_1) & 0 \\ 0 & (1-c)(1-g_0)+d(c+g_0-cg_0) & 0 & (1-d)(c+g_0-cg_0) \\ 0 & 1-(1-d)(1-(1-c)(1-g_0)(1-g_1)) & 0 & (1-d)(1-(1-c)(1-g_0)(1-g_1)) \end{pmatrix} \end{array}$$

L'estimation du modèle est simple. La vraisemblance est directement calculable car la banque de graines est toujours connue donc aucune variable n'est cachée. Les auteurs ont montré sur données simulées qu'il est possible de distinguer, en utilisant des méthodes de sélection de modèles, si une dynamique est avec ou sans dormance. Ils ont aussi prouvé que le modèle est capable de détecter la présence de graines dans la banque de graines.

Limites

Les populations sont modélisées par des données de type présence/absence et les hypothèses du modèle sont très restrictives. L'hypothèse que la dormance est limitée à 1 an n'est pas vérifiée pour toutes les espèces d'adventices. La colonisation est modélisée par pluie de graines.

2.2.4 Modélisation de la dynamique locale et régionale

Un modèle dans lequel la colonisation est représentée en pluie de graines ne permet pas de modéliser la dynamique régionale d'une espèce. L'influence de la colonisation des patchs voisins sur le patch d'arrivée décroît quand la distance entre la source et l'arrivée augmente.

2.2.4.1 Modèle de Levin et al

Le modèle de Levin et al. (1984) est utilisé pour étudier l'évolution de la stratégie de dispersion en prenant en compte la dormance. Pour cela, les auteurs étudient la stratégie évolutivement stable

de leur modèle. La dynamique des plantes est modélisée de manière déterministe à partir du nombre de graines dans la banque de graines du patch i, S_i avec $i \in \{1, ..., L\}$. L'évolution de la population est déterminée par :

$$S_i(t+1) = S_i(t)GY_i(t)(1-D) + S_i(t)(1-G)\nu + \frac{\alpha DG}{L}\sum_j S_j(t)Y_j(t)$$

où $Y_i(t)$ représente le nombre de graines produites, D est la fraction de dispersion et G est la fraction des graines qui vont germer. Le paramètre ν correspond au taux de survie des graines qui n'ont pas germé et α correspond à la proportion de graines qui ne sont pas perdues pendant la dispersion. Le nombre de graines produites est donné par :

$$Y_i(t) = K_i(t)F(GS_i(t))$$

où $K_i(t)$ est une variable aléatoire et F est une fonction qui prend en compte l'effet de la densité-dépendance des populations non dormantes. Ici, les auteurs ont choisi $F(x) = 1/x$.

Limites

Le modèle est déterministe et ne considère pas que la banque de graines est inconnue.

2.2.4.2 Modèle de Venable et al

Le modèle de Venable D and Brown (1988) prend en compte l'influence des populations voisines dans la dynamique de l'espèce. Il est utilisé pour comprendre comment la colonisation, la taille des graines et la dormance des graines évoluent dans le but de réduire le risque d'extinction en fonction des facteurs environnementaux.

Les auteurs modélisent les plantes annuelles avec dormance à partir de leur croissance par année :

$$\lambda = [1 - D(1-a)]\sum_{j=1}^{n}\rho_j(GS_j + R(1-G))$$

avec

- $GS_{ij} + R(1-G)$ est la production de graines du patch j au pas de temps suivant,
- D est la fraction de graines produites par une plante qui sont dispersées,
- G est la fraction de graines qui germent,
- R est la fraction de graines qui survivent dans la banque de graines pendant une année,
- a est le taux de survie des graines pendant la dispersion,
- n est le nombre de patchs,
- ρ_j est la probabilité qu'il y ait au moins une graine dans le sol du patch j ,
- S_j correspond à la production de graines par plante adulte du patch j.

Limites

La dynamique des espèces avec stade de dormance est modélisée à partir des populations dormantes, ce qui rend ce modèle inutilisable si seule l'information sur les populations non dormantes est disponible.

2.2.4.3 Modèle de Amarasekare et Possingham

Le modèle de Amarasekare and Possingham (2001) étudie la colonisation, l'extinction, la perturbation et la rapidité de succession dans les patchs influant sur l'équilibre de la population. Ce modèle présente l'originalité de tenir compte des conditions environnementales. On suppose ainsi qu'en saison défavorable, l'espèce n'est visible dans aucun patch.

On note

- I la proportion de patchs occupés avec des conditions environnementales favorables,
- S la proportion de patchs non occupés avec des conditions environnementales favorables,

— L la proportion de patchs occupés avec des conditions environnementales défavorables,
— R la proportion de patchs non occupés avec des conditions environnementales défavorables,
— g le taux de transition de patch favorable à défavorable,
— f le taux de transition de patch défavorable à favorable,
— e_I le taux d'extinction de la population dans des patchs favorables,
— e_L le taux d'extinction de la population dans des patchs défavorables,
— β_I le taux de colonisation dans des patchs favorables,
— β_L le taux de colonisation dans des patchs défavorables.

La transition de conditions environnementales est indépendante de l'état de la population dans le patch. La figure 2.7 représente les transitions possibles.

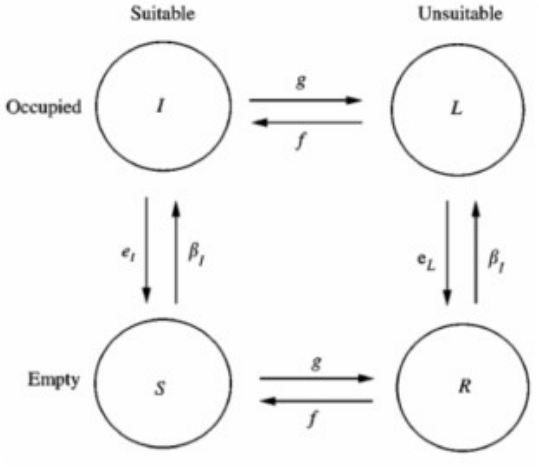


FIGURE 2.7 – Fgure extraite de Amarasekare and Possingham (2001)

Ainsi le système d'équations différentielles associé à la dynamique des 4 types de patchs est

$$\begin{aligned}
\frac{dI}{dt} &= \beta_I SI - e_I + fL - gI \\
\frac{dS}{dt} &= e_I - \beta_I SI + fR - gS \\
\frac{dL}{dt} &= gI - fL - e_L L + \beta_L RI \\
\frac{dR}{dt} &= gS - fR + e_L L - \beta_L RI
\end{aligned}$$

Limites

L'espèce est modélisée par des données de type présence/absence. Le modèle a besoin de connaître les saisons adéquates et les saisons inadéquates pour l'espèce.

2.2.4.4 Modèle de Mistro et al

Mistro et al. (2005) étudient l'équilibre de leur modèle à l'aide de simulations sans proposer de méthode d'estimation. Leur modèle suppose que la dormance des graines est limitée à 1 an. s_t représente le nombre de graines qui germent de la génération t et p_t représente le nombre de plantes adultes de la génération t.

Le nombre de graines qui germent peut être exprimé de telle sorte :

$$s_{t+1} = \alpha\sigma\rho p_t + \beta\sigma^2(1-\alpha)\rho p_{t-1}$$

où α, β et σ représentent respectivement la fraction des graines qui germent dès la première année, la fraction de celles qui germent la deuxième année et la fraction de celles qui survivent un hiver. ρ est le nombre de graines produites par plante. Le nombre de plantes adultes peut être exprimé par

$$p_{t+1} = as_{t+1}\,\mathrm{e}^{-bs_{t+1}}$$

où a représente la fécondité par graine qui germe et b représente le paramètre de densité-dépendance.

Afin d'introduire une variation spatiale, les auteurs considèrent que le nombre de graines qui germent et le nombre de plantes adultes dépendent de leur position géographique x. Pour ce faire, ils introduisent S_t la densité de graines qui germent au temps t et P_t la densité de plantes adultes au temps t. Ainsi on obtient

$$\begin{aligned} S_{t+1}(x) &= \alpha\sigma\rho \int_\Omega k(x,y)P_t(y)dy + \beta\sigma(1-\alpha)\sigma\rho \int_\Omega P_{t-1}(y)dy, \\ P_{t+1}(x) &= aS_{t+1}(x)\,\mathrm{e}^{-bS_{t+1}(x)} \end{aligned}$$

où Ω est l'habitat unidimensionnel et k est une fonction de redistribution, aussi connue sous le nom de noyau de dispersion, qui représente la probabilité qu'une graine d'une plante de la région $[y, y+dy)$ tombe dans x. Plus précisément, pour les simulations, k est défini par

$$k(x,y) = \frac{\mathrm{e}^{-|x-y|}}{2}.$$

Le modèle présente de nombreux points positifs. L'interaction entre les patchs dépend de leur distance. De plus, la croissance des plantes dépend de la densité des graines qui germent.

Limites

Le modèle considère que la dormance des graines ne peut dépasser 2 ans.

2.2.4.5 Modèle de Garnier et al

Le modèle de Garnier et al. (2006) a été créé afin d'étudier la stratégie optimale de survie de la dynamique locale et spatiale du colza. Ce modèle n'est pas destiné à de l'estimation. Les paramètres ont été fixés selon des données obtenues dans la littérature et selon des études menées sur le terrain. Le but est de déterminer les processus clefs qui assurent la persistance de la population. Les auteurs ont montré qu'augmenter la fréquence d'immigration avait un effet positif sur la persistance de la banque de graines.

Les auteurs utilisent l'information sur la survie des plantules ainsi que sur les méthodes de désherbage employées par les agriculteurs. Le modèle repose sur les paramètres suivants :

— J_{crop} la probabilité de colonisation venant des cultures de colza voisines,
— J_{transp} la probabilité de colonisation par pluie de graines,
— σ_{Rdd} la densité-dépendance de la survie des plantules,
— σ_0 le taux de survie des plantes adultes non-désherbées,
— σ_1 le taux de survie des plantes adultes désherbées,
— f_{0dd} la densité-dépendance de la fécondité des plantes adultes non-désherbées,
— f_{1dd} la densité-dépendance de la fécondité des plantes adultes désherbées,
— e_{S0w} le taux de germination des jeunes graines,
— e_{Sw} le taux de germination des vieilles graines,
— $\bar{e_{S0}}$ le taux de survie des jeunes graines,
— $\bar{e_S}$ le taux de survie des vieilles graines.

Voici comment sont calculés le nombre de graines N_s, le nombre de rosettes qui peuvent survivre N_r et le nombre de plantes adultes avant potentiel désherbage :

$$N_s(t+1) \sim \mathrm{Bin}\left[\mathrm{Pois}(f_{0dd}Fr_0(t)) + \mathrm{Pois}(f_{1dd}Fr_1(t)) + J_{crop} + J_{transp}, \bar{e_{S0}}\right] + \mathrm{Bin}[N_s(t), \bar{e_S}]$$

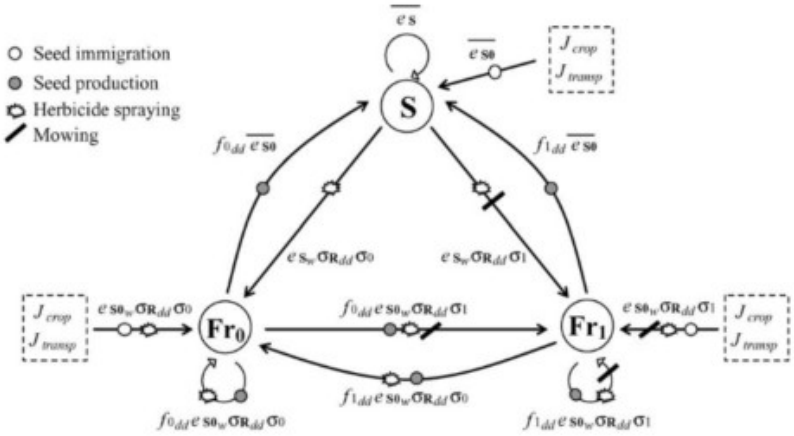


FIGURE 2.8 – figure extraite de Garnier et al. (2006)

$$N_r(t{+}1) \sim \text{Bin}(\text{Bin}\left[\text{Pois}(f_{0dd}Fr_0(t)) + \text{Pois}(f_{1dd}Fr_1(t)) + J_{crop} + J_{transp}, e_{S0w}\right] + \text{Bin}\left[N_s(t), e_{Sw}\right], \sigma_{Rdd})$$

où Bin est une loi Binomiale et Pois est une loi de Poisson.

Limites

Comme souligné précédemment, ce modèle n'a pas été créé dans le but de faire de l'estimation. Si malgré tout on souhaite l'utiliser dans cette perspective, le nombre important de paramètres sur lequel il repose le rendrait très probablement non identifiable.

2.2.4.6 Modèle de Manna et al

Le modèle de Manna et al. (2017) est une extension spatiale du modèle de Fréville et al. (2013). Comme dans ce dernier, les populations sont modélisées par présence/absence. Les populations de plantes sont observables tandis que celles de graines ne le sont pas. Le modèle a 4 paramètres. Les paramètres sont identiques au modèle de Fréville et al à l'exception de la colonisation. La probabilité de colonisation dépend des patchs voisins et n'est pas modélisée sous la forme d'une pluie de graines. Le modèle considère que les patchs sont arrangés en cercle et que la colonisation ne s'effectue qu'entre deux patchs côte à côte (figure 2.9 (B)). La probabilité de colonisation venant des patchs voisins est

$$P_c(Q) = 1 - (1 - c)^Q$$

ou $Q \in \{0, 1, 2\}$ est le nombre de patchs voisins avec une population non nulle de plantes et c représente la probabilité qu'un patch présentant des plantes colonise un patch voisin.

Les définitions et hypothèses sont identiques au modèle de Fréville et al. Ainsi, le modèle est une chaîne de Markov. La matrice de transition d'un patch est définie par

$$\begin{array}{c} \\ {}^{A}_{A} \\ {}^{A}_{P} \\ {}^{P}_{A} \\ {}^{P}_{P} \end{array} \begin{array}{c} \begin{array}{cccc} {}^{A}_{A} & {}^{A}_{P} & {}^{P}_{A} & {}^{P}_{P} \end{array} \\ \begin{pmatrix} 1 - P_c(1-d) & 0 & P_c(1-d) & 0 \\ (1-P_c)(1-g_1) + d(P_c + g_1 - P_c g_1) & 0 & (1-d)(P_c + g_1 - P_c g_1) & 0 \\ 0 & (1-P_c)(1-g_0) + d(P_c + g_0 - P_c g_0) & 0 & (1-d)(P_c + g_0 - P_c g_0) \\ 0 & 1 - (1-d)(1-(1-P_c)(1-g_0)(1-g_1)) & 0 & (1-d)(1-(1-P_c)(1-g_0)(1-g_1)) \end{pmatrix} \end{array}$$

L'estimation du modèle est simple car il suffit de maximiser la vraisemblance du modèle qui est calculable, la banque de graines étant toujours connue.

A partir de ce modèle, il est possible de définir des sous-modèles dans lesquels il n'y aurait pas de colonisation ou pas de survie. Les auteurs ont montré sur données simulées qu'il est possible, en utilisant des méthodes de sélection de modèles, de déterminer s'il y a connectivité entre les patchs et si la dynamique de l'espèce utilise un stade de dormance.

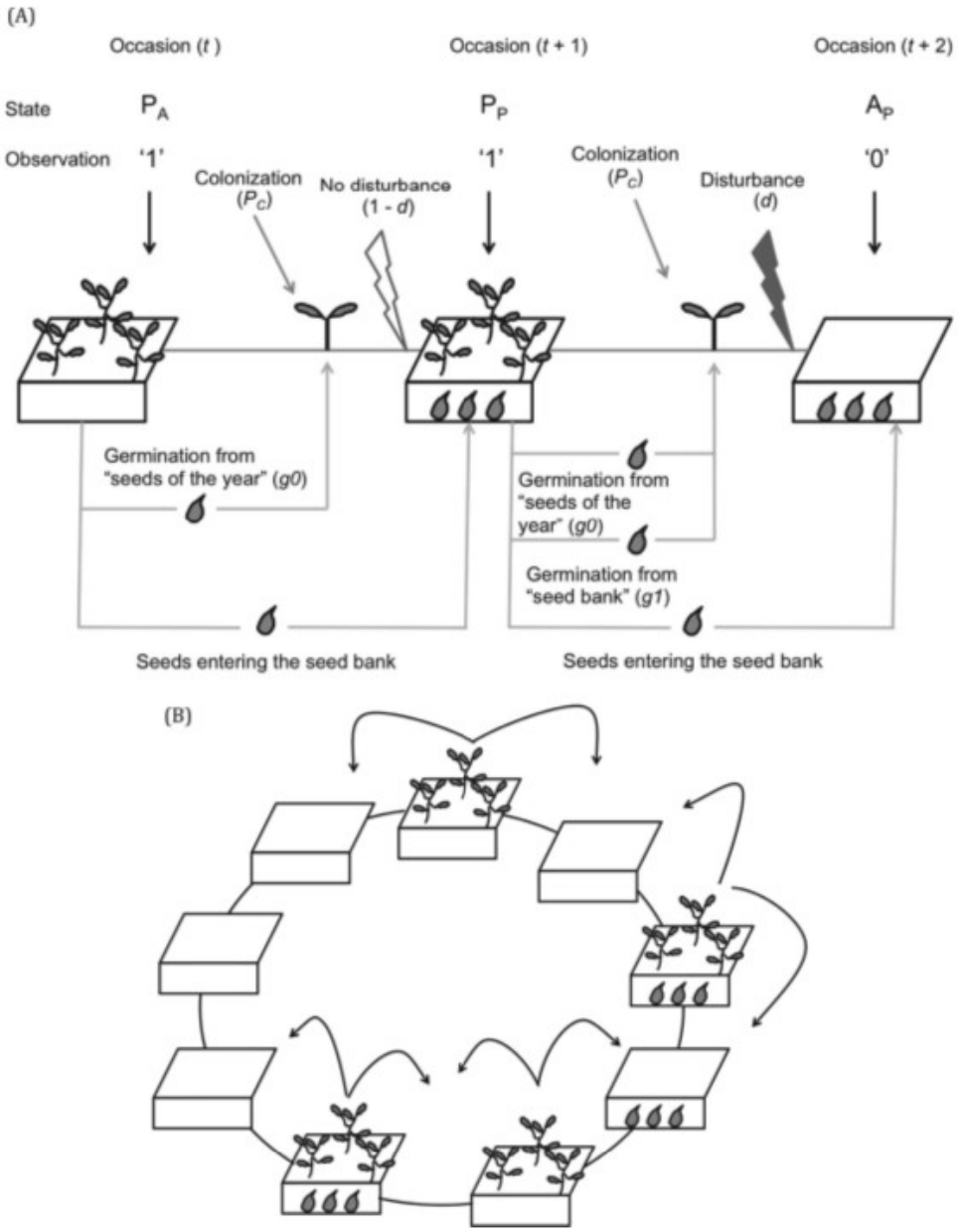


FIGURE 2.9 – Figure extraite de Manna et al. (2017)

Limites

Les limites de ce modèle sont identiques à celles du modèle de Fréville et al. à l'exception de la colonisation. Cependant la colonisation entre patchs ne fait intervenir que les deux plus proches voisins.

2.2.5 Discussion

Au regard des qualités visées pour le modèle que nous souhaitons établir, les modèles exposés ci-dessus présentent tous au moins un défaut majeur.

Pour représenter la dynamique régionale de l'espèce, il est nécessaire de prendre en compte la colonisation entre patchs. Ainsi les modèles qui ne font pas intervenir de colonisation ou ne la supposent que par pluie de graines ne modélisent qu'une dynamique locale.

Les modèles décrits qui intègrent la colonisation entre patchs, à l'exception de Manna et al. (2017), ne sont pas créés à des fins d'estimation mais visent en général à modéliser le plus exactement possible la réalité écologique. Il est raisonnable de penser que plus le nombre de paramètres est grand, meilleure sera la modélisation car les processus sont généralement modélisés de façon plus détaillée. Rappelons que notre objectif est de déterminer quels paramètres régissent la dynamique de l'espèce à partir des données sur la population observable. Plus un modèle prend en compte de paramètres, plus on a de risques d'avoir des jeux de paramètres différents qui engendrent des observations identiques, ce qui rendrait un tel modèle non identifiable et donc inutilisable à des fins d'estimation.

Quand bien même on souhaiterait utiliser ces modèles pour de l'estimation de la dynamique régionale, il faudrait en plus vérifier que l'état de la banque de graines soit bien considéré comme caché dans l'estimation. Ceci est crucial puisque dans la majorité des cas, seules les données sur la flore levée sont disponibles. Pour modéliser au mieux la dynamique, on s'intéresse aux processus écologiques, tels que la survie ou la dormance, susceptibles d'agir sur la banque de graines, en imposant le moins d'hypothèses possible sur ceux-ci. Une partie des modèles présentés, quand ils considèrent la dormance des graines, la supposent au plus de deux ans, après quoi la graine meurt. De plus , le modèle de Manna et al. (2017) suppose que s'il y a flore levée à un temps donné, la banque de graines sera présente au temps suivant. Au final, ces hypothèses restrictives sur les interactions avec la banque de graines impliquent que l'état de celle-ci, même s'il est caché à l'état initial, finit par être connu.

Enfin, reste le problème sur la nature des données. Les données sur les populations dans les modèles présentés sont majoritairement soient en termes de présence/absence, soit en termes d'effectifs. Dans la section 4.2 du chapitre 4 nous expliquons pourquoi nous utiliserons des classes d'abondance pour modéliser la dynamique des adventices.

2.3 Chaînes de Markov cachée

Dans la suite, nous présentons une approche pour estimer la dynamique locale et régionale d'une espèce à l'aide d'un modèle génériquement identifiable sans restriction sur l'âge de la banque de graines et à partir de données en classes d'abondance. Pour cela, nous avons opté pour une approche markovienne avec état caché où la population d'une espèce dépend de son état au temps précédent. Afin de comprendre le fonctionnement du modèle markovien, cette section est dédiée au rappel des définitions et propriétés des chaînes de Markov et des chaînes de Markov cachées ainsi qu'à la description des algorithmes classiques pour leur estimation : Expectation maximisation (Dempster et al., 1977) (EM) et l'algorithme de Monté Carlo (MCMC) (Ghahramani, 1998). Les notations utilisées dans la suite du document sont définies dans le chapitre 9 à la p140.

Les Chaînes de Markov constituent un outil mathématique largement utilisé dans de nombreux domaines. Elles sont employées, par exemple, pour mesurer l'indice de popularité d'une page Web (PageRank Rai and Lal (2016)), pour prédire le cours de la bourse Gupta and Dhingra (2012) ou encore pour la reconnaissance vocale (Rabiner, 1989). Afin de représenter une dynamique d'individus non observables, on dispose d'une variante des chaînes de Markov : les chaînes de Markov cachées. Elles ont déjà été utilisées dans des domaines varier comme pour le contrôle d'information visuelle (Rimey and Brown, 1991) ou bien en écologie pour des modèles de mouvement animal (Patterson et al., 2009), de capture-marquage (Patterson et al., 2008) ainsi que pour modéliser la succession écologique d'espèces (Usher, 1981).

2.3.1 Chaînes de Markov

Définition 1 (Chaîne de Markov discrète)**.** Une *chaîne de Markov discrète* est un processus stochastique à temps discret, c'est-à-dire une suite de v.a. $(X_n)_{n\in\mathbb{N}^*}.()$, définies sur un espace probabilisé $(\Omega, F, \mathbb{P})$, à valeurs dans un espace Ω_X appelé **espace d'états** (Ω_X sera supposé dénombrable ou fini), telle que pour tout entier $n \geq 0$ et pour tout n-uplets d'états $(x_0, ..., x_n) \in \Omega_X^n$,

$$\mathbb{P}(X_{n+1} = x_{n+1}|X_0 = x_0, ..., X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n).$$

Ainsi l'état de la variable aléatoire X_{n+1} ne dépend que de l'état de son prédécesseur X_n. On note $A_n(x_n, x_{n+1}) = \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n)$.

Une chaîne de Markov peut être définie à l'aide de deux probabilités de transition, la probabilité initiale et la probabilité de transition. Voici leur définition ci-dessous.

Définition 2 (Matrice de transition)**.** Pour tout entier $n \in \mathbb{N}^*$, on définit $A_n = ((A_n(x_{n-1}, x_n))_{x_{n-1},x_n\in\Omega_X}$ *la matrice de transition à l'instant* n.

Propriété 3. *Une matrice de transition* $A = ((A(x_n, x_{n+1}))_{x_n,x_{n+1}\in\Omega_X}$ *est stochastique, c'est-à-dire qu'elle vérifie les deux propriétés suivantes :*

1. $\forall\, x_n, x_{n+1} \in \Omega_X,\ A(x_n, x_{n+1}) \geq 0$,
2. $\forall\, x \in \Omega_X,\ \sum_{x_{n+1}\in\Omega_X} A(x_n, x_{n+1}) = 1$.

Définition 4 (Loi initiale de la chaîne)**.** La mesure de probabilité π sur Ω_X définie pour tout $x \in \Omega_X$ par $\pi(x) = \mathbb{P}(X_0 = x)$ est appelée *la loi initiale de la chaîne.*

Définition 5 (Chaîne Homogène)**.** La chaîne de Markov $(X_n)_{n\in\mathbb{N}^*}$ est dite *homogène* si la suite (A_n) est constante. On note alors $A_n = A$ pour tout $n \geq 1$.

Dorénavant, sauf mention contraire, la chaîne est supposée homogène.

Propriété 6. *Le couple* (π, A) *caractérise la loi de la chaîne de Markov* $(X_n)_{n\in\mathbb{N}^*}$ *si et seulement si*

$\forall N \geq 1$, $\forall x_0, ..., x_N \in \Omega_X$,

$$\mathbb{P}(X_0 = x_0, ..., X_N = x_N) = \pi(x_0) \prod_{n=1}^{N} A(x_{n-1}, x_n).$$

On dit alors que $(X_n)_{n\in\mathbb{N}^*}$ est une chaîne de Markov homogène de loi initiale π et de matrice de transition A.

2.3.2 Chaîne de Markov Cachée

Définition 7 (Chaîne de Markov cachée (HMM))**.** Une chaîne de Markov cachée est un processus stochastique double (X, Y) tel que

— $X = (X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov homogène,

— $Y = (Y_n)_{n\in\mathbb{N}}$ est un processus observable à valeurs dans Ω_Y indépendant conditionnellement à X.

Pour $N \in \mathbb{N}$, on pose $Y^N = (Y_0, ..., Y_N)$, $y^N = (y_0, ..., y_N)$, $X^N = (X_0, ..., X_N)$ et $x^N = (x_0, ..., x_N)$. L'hypothèse d'indépendance conditionnelle de Y sachant la chaîne de Markov X s'écrit

$$\forall\, N \in \mathbb{N}, \forall\, Y^N \in \Omega_Y^N,\ \forall\, X^N \in \Omega_X^N,\ \mathbb{P}(Y^N = y^N | X^N = x^N) = \prod_{n=0}^{N} \mathbb{P}(Y_n = y_n | X_n = x_n).$$

Ainsi il est possible de définir la loi d'émission de la chaîne qui représente la probabilité des variables observées conditionnellement aux variables cachées.

Définition 8 (Loi d'émission de la chaîne)**.** On définit la matrice des probabilités d'émission de l'observation Y_n sachant X_n, notée ϕ_n, par

$$\forall\, y \in \Omega_Y,\ \forall\, x \in \Omega_X,\ \phi_n(x, y) = \mathbb{P}(Y_n = y | X_n = x)$$

Par la suite on supposera que la matrice d'émission est homogène. Par conséquent, ϕ_n ne dépendant pas du temps n, c'est-à-dire que pour tout $n \geq 0$,

$$\forall\, y \in \Omega_Y,\ \forall\, x \in \Omega_X,\ \phi_n(x, y) = \mathbb{P}(Y_n = y | X_n = x) = \mathbb{P}(Y_0 = y | X_0 = x) = \phi(x, y).$$

.

Propriété 9. *Le triplet (π, A, ϕ) caractérise la chaîne de Markov cachée $((X_n), (Y_n))_{n\geq 0}$ si et seulement si*

$\forall n \geq 1$, $\forall x_0, ..., x_n \in \Omega_X$,

$$\mathbb{P}(X_0 = x_0, ..., X_N = x_N, Y_0 = y_0, ..., Y_N = y_n) = \pi(x_0)\phi(x_0, y_0) \prod_{n=1}^{N} A(x_{n-1}, x_n)\phi(x_n, y_n).$$

On dit alors que $((X_n), (Y_n))_{n\geq 0}$ est une chaîne de Markov cachée de loi initiale π, de loi d'émission ϕ et de probabilité de transition A.

La figure 2.10 correspond à la représentation graphique d'une chaîne de Markov cachée où les flèches représentent les dépendances entre les variables.

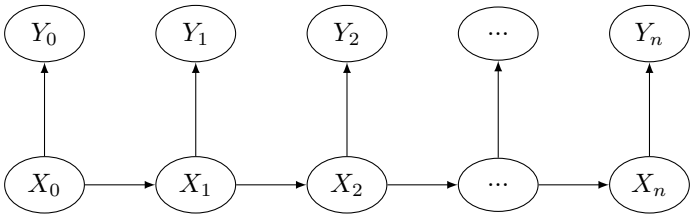


FIGURE 2.10 – Graphe d'une chaîne de Markov cachée

2.3.3 Estimation

L'estimation par maximum de vraisemblance permet, à partir de données, de trouver les paramètres les plus probables de générer ces données. En d'autres termes, l'estimation cherche à trouver les paramètres pour lesquels la probabilité d'observer ces données est la plus grande. Dans cette section, nous allons décrire comment calculer les maxima de vraisemblance pour des chaînes de Markov cachées.

2.3.3.1 Estimation des paramètres par maximum de vraisemblance

Dans le cas d'une chaîne de Markov simple, l'objectif est de maximiser la vraisemblance des observations par rapport à π et A. On cherche $(\hat{\pi}^{MV}, \hat{A}^{MV})$ tels que

$$(\hat{\pi}^{MV}, \hat{A}^{MV}) = \underset{(\pi,A)}{argmax}\ \mathbb{P}(X^N = x^N|\pi, A).$$

Dans un modèle de chaîne de Markov cachée, le traitement statistique est plus complexe car on n'observe pas la chaîne de Markov mais une suite émise par celle-ci. Par conséquent, la vraisemblance des observations nécessite d'être maximisée par rapport à π, A et ϕ. On note $\lambda = (\pi, A, \phi)$.

$$\hat{\lambda}^{MV} = \underset{\lambda}{argmax}\ \mathbb{P}(Y^N = y^N|\lambda)$$

Une maximisation directe de la vraisemblance mène à une expression compliquée à résoudre. On pourrait utiliser un algorithme de descente de gradient pour trouver l'argument du maximum. Dans ce cas, il est crucial d'initialiser des paramètres dans un voisinage proche de la solution, sans quoi l'amélioration des estimateurs à chaque itération est mauvaise. Sachant que l'espace d'états risque d'être grand pour le modèle associé à la dynamique des adventices, si l'on n'a pas d'*a priori* sur la dynamique, l'initialisation des estimateurs sera aléatoire. Même si la convergence s'améliore grandement une fois dans un voisinage correct, nous ne comptons pas utiliser l'algorithme de descente de gradient pour maximiser la vraisemblance.

On choisit d'utiliser l'algorithme EM (Dempster et al., 1977) qui présente une approche différente. L'algorithme EM permet d'améliorer fortement la valeur des estimateurs dès la première itération. Cependant, il converge lentement une fois les estimateurs dans le voisinage des vrais paramètres.

2.3.3.2 Le principe de l'algorithme EM

L'algorithme EM (Dempster et al., 1977) estime de façon itérative les paramètres du modèle. Pour cela, il utilise la vraisemblance dans le cas de données complètes (x^N, y^N). Si x^N était observé, on chercherait à la maximiser par rapport à λ.

$$\hat{\lambda}^{MVC} = \underset{\lambda}{argmax}\ \mathbb{P}(Y^N = y^N, X^N = x^N|\lambda)$$

Étant donné que x^N n'est pas disponible, on choisit de maximiser l'espérance de la log vraisemblance complète conditionnellement aux données observées et à λ. Notons λ_{it-1} la valeur de l'estimateur à l'itération $it-1$. L'algorithme EM itère 2 étapes :

1. Étape E : Calcul des probabilités nécessaires au calcul de l'espérance de la log vraisemblance complète conditionnellement aux données observées et à λ_{it-1}.
2. Étape M : Estimation du paramètre λ_{it} avec

$$\lambda_{it} = \arg\max_{\lambda} \mathbb{E}[\ln(\mathbb{P}(X^N, Y^N|\lambda))|Y^N = y^N, \lambda_{it-1}]$$

L'algorithme EM garantit que l'estimateur $\hat{\lambda}^{MV}$ du maximum vraisemblance des observations est un point fixe de l'algorithme EM, autrement dit

$$\lambda^{MV} = \arg\max_{\lambda} \mathbb{E}[\ln(\mathbb{P}(X^N, Y^N|\lambda))|Y^N = y^N, \lambda^{MV}].$$

En effet, la log-vraisemblance et la log-vraisemblance complète vérifient la relation :

$$\ln(\mathbb{P}(X^N = x^N, Y^N = y^N|\lambda) = \ln(\mathbb{P}(X^N = x^N|Y^N = y^N, \lambda) + \ln(\mathbb{P}(Y^N = y^N|\lambda).$$

Prenons l'espérance de la relation précédente conditionnellement aux données observées et à λ^{MV}. Ainsi

$$\begin{aligned}\mathbb{E}[\ln(\mathbb{P}(X^N, Y^N|\lambda)|Y^N = y^N, \lambda^{MV})] &= \mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda))|Y^N = y^N, \lambda^{MV})] \\ &\quad +\mathbb{E}[\ln(\mathbb{P}(Y^N|\lambda))|Y^N = y^N, \lambda^{MV})] \\ &= \mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda))|Y^N = y^N, \lambda^{MV})] \\ &\quad + \ln(\mathbb{P}(Y^N = y^N|\lambda))\end{aligned}$$

Notons

$$\begin{aligned}\mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda))|Y^N = y^N, \lambda^{MV})] &= A(\lambda), \\ \ln(\mathbb{P}(Y^N = y^N|\lambda)) &= B(\lambda).\end{aligned}$$

Par définition on sait que pour $B(\lambda)$, le maximum est atteint en λ^{MV}.

Il nous suffit de montrer que $\lambda^{MV} = \arg\max_{\lambda} A(\lambda)$. Puisque ln est une fonction concave, il est possible d'appliquer l'inégalité de Jensen :

$$\mathbb{E}[\ln(\tfrac{\mathbb{P}(X^N|Y^N=y^N,\lambda)}{\mathbb{P}(X^N|Y^N=y^N,\lambda^{MV})})|Y^N = y^N, \lambda^{MV}) \leq \ln(\mathbb{E}[\tfrac{\mathbb{P}(X^N|Y^N=y^N,\lambda)}{\mathbb{P}(X^N|Y^N=y^N,\lambda^{MV})}|Y^N = y^N, \lambda^{MV}])$$

Cette inégalité est vérifiée si $\lambda = \lambda^{MV}$. Par ailleurs,

$$\begin{aligned}\ln(\mathbb{E}[\tfrac{\mathbb{P}(X^N|Y^N=y^N,\lambda)}{\mathbb{P}(X^N|Y^N=y^N,\lambda^{MV})}|Y^N = y^N, \lambda^{MV})) &= \ln\left(\sum_{x^N} 1 \times \mathbb{P}(X^N = x^N|Y^N = y^N, \lambda)\right) \\ &= \ln(1) \\ &= 0\end{aligned}$$

De plus,

$$\begin{aligned}\mathbb{E}[\ln(\tfrac{\mathbb{P}(X^N|Y^N=y^N,\lambda)}{\mathbb{P}(X^N|Y^N=y^N,\lambda^{MV})})|Y^N = y^N, \lambda^{MV}) &= \mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda))|Y^N = y^N, \lambda^{MV}) \\ &\quad -\mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda^{MV})|Y^N = y^N, \lambda^{MV})\end{aligned}$$

Pour tout λ, on obtient alors

$$\begin{aligned} A(\lambda) &= \mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda))|Y^N = y^N, \lambda^{MV}) \\ &\leq \mathbb{E}[\ln(\mathbb{P}(X^N|Y^N = y^N, \lambda^{MV})|Y^N = y^N, \lambda^{MV}) = A(\lambda^{MV}) \end{aligned}$$

Donc le maximum de A est atteint quand $\lambda = \lambda^{MV}$. Puisque A et $\mathcal{B}$ atteignent leur maximum quand $\lambda = \lambda^{MV}$, λ^{MV} est un point fixe de l'algorithme EM.

Il est important de noter que l'algorithme EM converge vers un maximum local. Ainsi, il est préférable de lancer l'algorithme plusieurs fois avec des valeurs d'initialisation différentes pour atteindre un maximum global.

2.3.3.3 Algorithme EM

A présent, détaillons les étapes de l'algorithme EM (Rabiner, 1989 ; Rabiner and Juang, 1986). D'abord, exprimons l'espérance conditionnelle de la log vraisemblance complète, notée LLE, en fonction de $\lambda = (\pi, A, \phi)$.

$$\begin{aligned} LLE = \mathbb{E}[\ln(\mathbb{P}(X^N, Y^N|\lambda)|Y^N = y^N, \lambda_{it}) =& \mathbb{E}[\ln(\mathbb{P}(Y^N|X^N = x^N, \lambda))|Y^N = y^N, \lambda_{it}) \\ &+ \mathbb{E}[\ln(\mathbb{P}(X^N|\lambda))|Y^N = y^N, \lambda_{it}) \end{aligned}$$

D'après la propriété 9, on peut réécrire l'espérance de la log vraisemblance sachant les données comme ci-après.

$$\begin{aligned} LLE =& \mathbb{E}[\ln(\prod_{n=0}^{N} \mathbb{P}(Y_n|X_n = x_n, \lambda))|Y^N = y^N, \lambda_{it}] \\ &+ \mathbb{E}[\ln(\pi(X_0) \times \prod_{n=1}^{N} A(x_{n-1}, X_n))|Y^N = y^N, \lambda_{it}] \\ =& \sum_{n=0}^{N} \mathbb{E}[\ln(\mathbb{P}(Y_n|X_n = x_n, \lambda))|Y^N = y^N, \lambda_{it}] \\ &+ \mathbb{E}[\ln(\pi(X_0))|Y^N = y^N, \lambda_{it}] \\ &+ \sum_{n=1}^{N} \mathbb{E}[\ln(A(x_{n-1}, X_n))|Y^N = y^N, \lambda_{it}] \\ =& \sum_{n=0}^{N} \sum_{x_n \in \Omega_X} \mathbb{P}(X_n = x_n|Y^N = y^N, \lambda_{it}) \ln(\phi_n(x_n, y_n)) \\ &+ \sum_{x_0 \in \Omega_X} \mathbb{P}(X_0 = x_0|Y^N = y^N, \lambda_{it}) \ln(\pi(x_0)) \\ &+ \sum_{n=1}^{N} \sum_{(x_{n-1}x_n) \in \Omega_X^2} \mathbb{P}(X_{n-1} = x_{n-1}, X_n = x_n|Y^N = y^N, \lambda_{it}) \ln(A(x_{n-1}, x_n)) \end{aligned}$$

Étape E :

Intéressons-nous maintenant à l'étape E qui correspond à l'algorithme du Forward-Backward (Rabiner (1989)) et réalise le calcul des probabilités conditionnelles nécessaires à l'étape M.

1. $\rho_n(x_n) = \mathbb{P}(X_n = x_n|Y^N = y^N, \lambda_{it})$
2. $\xi_n(x_{n-1}, x_n) = \mathbb{P}(X_{n-1} = x_{n-1}, X_n = x_n|Y^N = y^N, \lambda_{it})$

Pour cela utilisons les fonctions α_n et β_n que l'on peut facilement calculer pour tout n de manière récursive.

$$\alpha_n(x) = \mathbb{P}(Y^n = y^n, X_n = x|\lambda_{it}) \tag{2.1}$$
$$\beta_n(x) = \mathbb{P}(Y_{n+1} = y_{n+1}, ..., Y_N = y_N|X_n = x, \lambda_{it}) \tag{2.2}$$

On note $\lambda_{it} = (\pi_{it}, \phi_{it}, A_{it})$. Voici la relation de récurrence pour α_n, $\forall\, 0 \leq n \leq N$:

$$\begin{aligned}
\alpha_n(x_n) =& \mathbb{P}(Y^n = y^n, X_n = x_n|\lambda_{it}) \\
=& \mathbb{P}(Y^n = y^n|X_n = x_n, \lambda_{it}) \times \mathbb{P}(X_n = x_n|\lambda_{it}) \\
=& \mathbb{P}(Y^{n-1} = y^{n-1}|X_n = x_n, \lambda_{it}) \times \mathbb{P}(Y_n = y_n|X_n = x_n, \lambda_{it}) \times \mathbb{P}(X_n = x_n|\lambda_{it}) \\
=& \mathbb{P}(Y^{n-1} = y^{n-1}, X_n = x_n|\lambda_{it}) \times \mathbb{P}(Y_n = y_n|X_n = x_n, \lambda_{it}) \\
=& \mathbb{P}(Y_n = y_n|X_n = x_n, \lambda_{it}) \times \sum_{x_{n-1} \in \Omega_X} \mathbb{P}(Y^{n-1} = y^{n-1}, X_{n-1} = x_{n-1}, X_n = x_n|\lambda_{it}) \\
=& \phi_{it}(x_n, y_n) \sum_{x_{n-1} \in \Omega_X} \mathbb{P}(X_n = x_n|Y^{n-1} = y^{n-1}, X_{n-1} = x_{n-1}, \lambda_{it}) \mathbb{P}(Y^{n-1} = y^{n-1}, X_{n-1} = x_{n-1}|\lambda_{it}) \\
=& \phi_{it}(x_n, y_n) \times \sum_{x_{n-1} \in \Omega_X} \alpha_{n-1}(x_{n-1}) \times A_{it}(x_{n-1}, x_n)
\end{aligned}$$

avec $\alpha_0(x_0) = \phi_{it}(x_0, y_0) \times \pi_{it}(x_0)$. Passons maintenant à β_n.
Ainsi de la même manière, voici la relation de récurrence pour β_n : $\forall\, 0 \leq n \leq N$,

$$\begin{aligned}
\beta_n(x_n) =& \mathbb{P}(Y_{n+1} = y_{n+1}, ..., Y_N = y_n|X_n = x_n, \lambda_{it}) \\
=& \sum_{x_{n+1} \in \Omega_X} \mathbb{P}(Y_{n+1} = y_{n+1}, ..., Y_N = y_n, X_{n+1} = x_{n+1}|X_n = x_n, \lambda_{it}) \\
=& \sum_{x_{n+1} \in \Omega_X} \mathbb{P}(Y_{n+1} = y_{n+1}, ..., Y_N = y_n|X_{n+1} = x_{n+1}, \lambda_{it}) \times \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n, \lambda_{it}) \\
=& \sum_{x_{n+1} \in \Omega_X} \mathbb{P}(Y_{n+2} = k_{n+2}, ..., Y_N = y_n|X_{n+1} = x_{n+1}, \lambda_{it}) \mathbb{P}(Y_{n+1} = y_{n+1}|X_{n+1} = x_{n+1}, \lambda_{it}) A_{it}(x_n, x_{n+1}) \\
=& \sum_{x_{n+1} \in \Omega_X} \beta_{n+1}(x_{n+1}) \times \phi_{it}(x_{n+1}, y_{n+1}) \times A_{it}(x_n, x_{n+1})
\end{aligned}$$

avec $\beta_N(x_N) = 1$.
De plus, α_n et β_n vérifient

$$\begin{aligned}
\alpha_n(x_n) \times \beta_n(x_n) =& \mathbb{P}(Y^n = y^n, X_n = x_n|\lambda_{it}) \times \mathbb{P}(Y_{n+1} = y_{n+1}, ..., Y_N = y_n|X_n = x_n, \lambda_{it}) \\
=& \mathbb{P}(X_n = x_n|\lambda_{it}) \times \mathbb{P}(Y^n = y^n|X_n = x_n, \lambda_{it}) \times \mathbb{P}(Y_{n+1} = y_{n+1}, ..., Y_N = y_n|X_n = x_n, \lambda_{it}) \\
=& \mathbb{P}(X_n = x_n|\lambda_{it}) \times \mathbb{P}(Y^N = y^N|X_n = x_n, \lambda_{it}) \\
=& \mathbb{P}(Y^N = y^N, X_n = x_n|\lambda_{it})
\end{aligned}$$

et

$$\sum_{x_n \in \Omega_X} \alpha_n(x_n) \times \beta_n(x_n) = \mathbb{P}(Y^N = y^N|\lambda_{it}). \tag{2.3}$$

Grâce à ces relations, il est possible de calculer $\rho_n(x_n)$ pour tout $x_n \in \Omega_X$ et $\xi_n(x_{n-1}, x_n)$ pour tout $(x_{n-1}, x_n) \in \Omega_X^2$.

$$\rho_n(x_n) = \mathbb{P}(X_n = x_n | Y^N = y^N, \lambda_{it}) = \frac{\alpha_n(x_n) \times \beta_n(x_n)}{\sum\limits_{x_n \in \Omega_X} [\alpha_n(x_n) \times \beta_n(x_n)]}$$

$$\begin{aligned}\xi_n(x_{n-1}, x_n) =& \mathbb{P}(X_{n-1} = x_{n-1}, X_n = x_n | Y^N = y^N, \lambda_{it}) \\ =& \frac{\mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1} | \lambda_{it}) \times \mathbb{P}(Y^N = y^N | X_n = x_n, X_{n-1} = x_{n-1}, \lambda_{it})}{\mathbb{P}(Y^N = y^N | \lambda_{it})}\end{aligned}$$

Or $\mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1} | \lambda_{it}) = A_{it}(x_{n-1}, x_n)\mathbb{P}(X_{n-1} = x_{n-1} | \lambda_{it})$ et

$$\begin{aligned}\mathbb{P}(Y^N = y^N | X_n = x_n, X_{n-1} = x_{n-1}, \lambda_{it}) =& \mathbb{P}(Y^{n-1} = y^{n-1} | X_{n-1} = x_{n-1}, \lambda_{it}) \phi_{it}(x_n, y_n) \\ & \times \mathbb{P}(Y_{n+1} = y_{n+1}, .., Y_N = y_n | X_n = x_n, \lambda_{it})\end{aligned}$$

En utilisant (2.1), (2.2) et (2.3), on a alors

$$\xi_n(x_{n-1}, x_n) = \frac{A_{it}(x_{n-1}, x_n) \times \alpha_{n-1}(x_{n-1}) \times \phi_{it}(x_n, y_n) \times \beta_n(x_n)}{\sum\limits_{x_n \in \Omega_X} [\alpha_n(x_n) \times \beta_n(x_n)]}$$

Étape M :

Les paramètres étant des probabilités, il y a des contraintes de normalité qui sont prises en compte par la méthode des multiplicateurs de Lagrange. Ainsi, il est possible d'obtenir un estimateur λ_{it+1} en maximisant l'espérance de la log-vraisemblance du modèle conditionnellement aux données.

$$\begin{aligned}L(\pi, A, \eta_1, \eta_2, \eta_3) =& \mathbb{E}[\ln(\mathbb{P}(X^N, Y^N | \lambda)) | Y^N = y^N, \lambda_{it}] - \eta_1 \left[\left(\sum_{x_0 \in \Omega_X} \pi(x_0) \right) - 1 \right] \\ & - \eta_2 \left[\left(\sum_{x_n \in \Omega_X} A(x_{n-1}, x_n) \right) - 1 \right] - \eta_3 \left[\left(\sum_{y \in \Omega_Y} \phi(x, y) \right) - 1 \right]\end{aligned}$$

Commençons par déterminer les estimateurs de π et de A puis nous estimerons une probabilité d'émission dans le cas dépendant et indépendant du temps. Il est possible d'estimer indépendamment π, A et ϕ ou ϕ_n car la maximisation de Lagrange peut être vue comme 3 systèmes d'équations linéaires indépendantes entre eux. En effet, il y a un système d'équations linéaires pour chaque paramètre du modèle. On pose

$$\begin{aligned}\frac{\partial L(\pi, A, \eta_1, \eta_2)}{\partial \pi(x_0)} =& \frac{\mathbb{P}(X_0 = x_0 | Y^N = y^N, \lambda_{it})}{\pi(x_0)} - \eta_1 \\ \frac{\partial L(\pi, A, \eta_1, \eta_2)}{\partial A(x, x')} =& \left(\sum_{n=1}^{N} \frac{\mathbb{P}(X_{n-1} = x, X_n = x' | Y^N = y^N, \lambda_{it})}{A(x, x')} \right) - \eta_2 \\ \frac{\partial L(\pi, A, \eta_1, \eta_2)}{\partial \eta_1} =& \left(\sum_{x_0 \in \Omega_X} \pi(x_0) \right) - 1 \\ \frac{\partial L(\pi, A, \eta_1, \eta_2)}{\partial \eta_2} =& \left(\sum_{x' \in \Omega_X} A(x, x') \right) - 1\end{aligned}$$

Puisque la matrice de transition A ne dépend pas de n, elle peut être sortie de la somme de la deuxième équation. Quand les dérivées partielles sont nulles on obtient

$$
\begin{aligned}
\eta_1 \pi(x_0) &= \mathbb{P}(X_0 = x_0 | Y^N = y^N, \lambda_{it}) \\
\eta_2 A(x, x') &= \sum_{n=1}^{N} \mathbb{P}(X_{n-1} = x, X_n = x' | Y^N = y^N, \lambda_{it}) \\
\sum_{x_0 \in \Omega_X} \pi(x_0) &= 1 \\
\sum_{x' \in \Omega_X} A(x, x') &= 1
\end{aligned}
$$

Il suffit ensuite de résoudre le système linéaire suivant :

$$
\begin{aligned}
\eta_1 =& 1 \\
\eta_2 =& \sum_{x' \in \Omega_X} \sum_{n=1}^{N} \xi(x, x') \ , \ \forall \, (x, x') \in \Omega_X^2 \\
\pi_{it+1}(x_0) =& \rho_0(x_0) \ , \ \forall \, x_0 \in \Omega_X \\
A_{it+1}(x, x') =& \frac{\sum_{n=1}^{N} \xi(x, x')}{\sum_{x' \in \Omega_X} \sum_{n=1}^{N} \xi(x, x')} \ , \ \forall \, (x, x') \in \Omega_X^2
\end{aligned}
$$

Enfin, l'estimateur de ϕ, ne dépendant pas du temps, est obtenu de la manière suivante.

$$
L'(\phi, \eta_3) = \mathbb{E}[\ln(\mathbb{P}(X^N, Y^N | \lambda)) | Y^N = y^N, \lambda_{it}) - \eta_3 \left[\left(\sum_{y \in \Omega_Y} \phi(x, y) \right) - 1 \right]
$$

En regardant le lieu d'annulation de L', on obtient pour tout $x \in \Omega_X$

$$
\phi_{it+1}(x, y) = \frac{\sum_{n \in \{n, y = y_n\}} \mathbb{P}(X_n = x | Y^N = y^N, \lambda_{it})}{\sum_{y \in \Omega_Y} \sum_{n \in \{n, y = y_n\}} \mathbb{P}(X_n = x | Y^N = y^N, \lambda_{it})} = \frac{\sum_{n \in \{n, y = y_n\}} \rho_n(x)}{\sum_{y \in \Omega_Y} \sum_{n \in \{n, y = y_n\}} \rho_n(x)} = \frac{\sum_{n \in \{n, y = y_n\}} \rho_n(x)}{\sum_{n=0}^{N} \rho_n(x)}
$$

De la même manière, voici l'estimateur de ϕ_n qui dépend du temps :

1. Si $y = y_n$, alors :

$$
\phi_{it+1,n}(x, y) = \frac{\mathbb{P}(X_n = x | Y^N = y^N, \lambda_{it})}{\sum_{y \in \Omega_Y} \mathbb{P}(X_n = x | Y^N = y^N, \lambda_{it})} = \frac{\rho_n(x)}{\sum_{y \in \Omega_Y} \rho_n(x)}
$$

2. Si $y \neq y_n$, $\phi_{it+1,n}(x, y)$ ne peut être estimé.

Il est important de remarquer que dans le cas où l'estimateur de ϕ_n dépend du temps, le nombre de paramètres est beaucoup plus important, ce qui rend l'estimation plus lente et moins précise.

2.3.3.4 Pseudo-code du EM

Nous présentons maintenant la mise en oeuvre concrète de l'algorithme EM. Pour commencer il faut d'initialiser un λ_0. Il faut ensuite itérer sur les étapes E et M.

On initialise les paramètres à λ_0 et on fixe ϵ puis on itère sur les étapes E et M jusqu'à ce que $|\lambda_{it-1} - \lambda_{it}| < \epsilon$.

Étape E :

Calcul de α_n et β_n

1. $\alpha_n(x_n) = \phi_{it}(x_n, y_n) \times \sum_{x_{n-1} \in \Omega_X} \alpha_{n-1}(x_{n-1}) \times A_{it}(x_{n-1}, x_n)$
2. $\beta_n(x_n) = \sum_{x_{n+1} \in \Omega_X} \beta_{n+1}(x_{n+1}) \times \phi_{it}(x_{n+1}, y_{n+1}) \times A_{it}(x_n, x_{n+1})$

Calcul de ξ_n et ρ_n.

1. $\rho_n(x_n) = \frac{\alpha_n(x_n) \times \beta_n(x_n)}{\sum_{x_n \in \Omega_X} [\alpha_n(x_n) \times \beta_n(x_n)]}$ pour tout $x_n \in \Omega_X$.
2. $\xi_n(x_{n-1}, x_n) = \frac{A(x_{n-1}, x_n) \times \alpha_{n-1}(x_{n-1}) \times \phi_n(x_n, y_n) \times \beta_n(x_n)}{\sum_{x_n \in \Omega_X} [\alpha_n(x_n) \times \beta_n(x_n)]}$ pour tout $(x_{n-1}, x_n) \in \Omega_X^2$.

Étape M :
Mise à jour des paramètres π, A et ϕ_n ou ϕ.

1. $\pi_{it+1}(x_0) = \rho_0(x_0)$, $\forall\, x_0 \in \Omega_X$
2. $A_{it+1}(x_{n-1}, x_n) = \frac{\sum_{n=1}^{N} \xi_n(x_{n-1}, x_n)}{\sum_{x_n \in \Omega_X} \sum_{n=1}^{N} \xi_n(x_{n-1}, x_n)}$, pour tout $(x_{n-1}, x_n) \in \Omega_X^2$
3. $\phi_{it+1}(x, y) = \frac{\sum_{n \in \{n, y_n = y\}} \rho_n(x)}{\sum_{n=0}^{N} \rho_n(x)}$, pour tout $x \in \Omega_X$, pour tout $y \in \Omega_Y$
4. Si $y_n = y$, alors $\phi_{it+1,n}(x, y) = \frac{\rho_n(x)}{\sum_{y \in \Omega_Y} \rho_n(x)}$, pour tout $x \in \Omega_X$, pour tout $y \in \Omega_Y$.

2.3.3.5 Algorithme EM pour plusieurs HMM indépendant de même loi

Dans cette section, on présente l'estimation des paramètres de plusieurs chaînes de Markov cachées indépendantes ayant la même loi. Cette section servira à étendre l'algorithme EM pour un modèle avec plusieurs chaînes de Markov cachées non-indépendantes.

Les C chaînes de Markov cachées seront notées $(X_{c,n})_{n \in \mathbb{N}^*}$ pour tout $c \in \{1, ..., C\}$. On notera ainsi $(Y_{c,n})_{n \geq 0}$ la suite d'observations produites par la $c^{\text{ième}}$ chaîne de Markov cachée $(X_{c,n})_{n \geq 0}$. On utilise les notations définies dans le chapitre 9.

Étape E :

L'étape E s'effectue de la même manière que dans la section 2.3.3.3. Ainsi, on calcule $\alpha_{c,n}$ et $\beta_{c,n}$ pour chaque chaîne de Markov $c \in \{1, ..., C\}$. Ces nombres servent à calculer les probabilités $\rho_{c,n}$ et $\xi_{c,n}$.

1. $\alpha_{c,n}(x_n) = \phi_{it}(x_n, y_{c,n}) \times \sum_{x_{n-1} \in \Omega_X} \alpha_{c,n-1}(x_{n-1}) \times A_{it}(x_{n-1}, x_n)$
2. $\beta_{c,n}(x_n) = \sum_{x_{n+1} \in \Omega_X} \beta_{c,n+1}(x_{n+1}) \times \phi_{it}(x_{n+1}, y_{c,n+1}) \times A_{it}(x_n, x_{n+1})$

avec $\alpha_{c,0}(x_0) = \phi_{it}(x_0, y_{c,0}) \times \pi_{it}(x_0)$ et $\beta_{c,N}(x_N) = 1$.

Calcul des probabilités $\xi_{c,n}$ et $\rho_{c,n}$.

1. $\rho_{c,n}(x_n) = \frac{\alpha_{c,n}(x_n) \times \beta_{c,n}(x_n)}{\sum_{x_n \in \Omega_X} [\alpha_{c,n}(x_n) \times \beta_{c,n}(x_n)]}$ pour tout $x_n \in \Omega_X$.
2. $\xi_{c,n}(x_{n-1}, x_n) = \frac{\hat{A}(x_{n-1}, x_n) \times \alpha_{c,n-1}(x_{n-1}) \times \hat{\phi}_n(x_n, y_n) \times \beta_{c,n}(x_n)}{\sum_{x_n \in \Omega_X} [\alpha_{c,n}(x_n) \times \beta_{c,n}(x_n)]}$ pour tout $(x_{n-1}, x_n) \in \Omega_X^2$.

Étape M :

La mise à jour des paramètres se fait à partir de la maximisation de Lagrange :

$$\begin{aligned}
\pi_{it+1}(x_0) &= \frac{1}{C}\sum_{c=1}^{C}\rho_{c,0}(x_0)\ ,\ \forall\, x_0 \in \Omega_X \\
A_{it+1}(x_{n-1},x_n) &= \frac{\sum_{c=1}^{C}\sum_{n=1}^{N}\xi_{c,n}(x_{n-1},x_n)}{\sum_{c=1}^{C}\sum_{x_n\in\Omega_X}\sum_{n=1}^{N}\xi_{c,n}(x_{n-1},x_n)}\ ,\ \forall\,(x_{n-1},x_n)\in\Omega_X^2 \\
\phi_{it+1}(x,y) &= \frac{\sum_{c=1}^{C}\sum_{n\in\{n,y_{c,n}=y\}}\rho_{c,n}(x)}{\sum_{c=1}^{C}\sum_{n=0}^{N}\rho_{c,n}(x)}\ ,\ \forall\, x\in\Omega_X
\end{aligned}$$

2.3.3.6 L'échantillonneur de Gibbs

L'échantillonneur de Gibbs (Rydén (2008)) est souvent utilisé pour des statistiques inférentielles, plus particulièrement pour l'inférence bayésienne. Il utilise des nombres aléatoires au sein de l'algorithme et présente une alternative à l'algorithme EM. L'échantillonneur de Gibbs est un algorithme itératif qui estime les lois des paramètres à l'aide de loi *a priori*.

Puisque Ω_X est un espace discret, on suppose que $\Omega_X = \{1,...,|\Omega_X|\}$. De la même manière on suppose que $\Omega_Y = \{1,...,|\Omega_Y|\}$. Notons

$$\pi = (\pi_1,...,\pi_{|\Omega_X|}),\ \phi = (\phi(x,y))_{x\in\Omega_X,y\in\Omega_Y},\ A = (A(x_n,x_{n+1}))_{x_n,x_{n+1}\in\Omega_X^2} \text{ et } \lambda_{\Omega_X} = (\pi,A,\phi_{\Omega_X}).$$

Sans autre information, nous utiliserons des lois *a priori* classiques sur les paramètres π, A, ϕ. Ainsi, on suppose que π suit une loi de Dirichlet de supports $(p(1),...,p(|\Omega_X|)) \in [0,1]^{|\Omega_X|}$ et de paramètres $(\mathbf{1}^C) \in [0,1]^{|\Omega_X|}$. De la même manière, on suppose que les lignes de la matrice A suivent une loi de Dirichlet de supports $(a(x,1),...,a(x,|\Omega_X|)) \in [0,1]^{|\Omega_X|}$ et de paramètres $(\mathbf{1}^C) \in [0,1]^{|\Omega_X|}$ et les lignes de la matrice ϕ suivent une loi de Dirichlet de supports $(\chi(x,1),...,\chi(x,|\Omega_Y|)) \in [0,1]^{|\Omega_X|}$ et de paramètres $(\mathbf{1}^C) \in [0,1]^{|\Omega_X|}$. On définit les lois *a priori* comme des lois de Dirichlet car les lois *a posteriori* seront aussi des lois de Dirichlet.

Nous rappelons la densité de la loi de Dirichlet de paramètres $(\nu_1,...,\nu_{|\Omega_X|})$ et de supports $(z_1,...,z_{|\Omega_X|})$:

$$f(z_1,...,z_{|\Omega_X|};\nu_1,...,\nu_{|\Omega_X|}) = \frac{\prod_{x=1}^{|\Omega_X|} z_x^{\nu_x-1}}{B(\nu)}$$

avec

$$B(\nu) = \frac{\prod_{x=1}^{|\Omega_X|}\rho(\nu_x)}{\rho\left(\sum_{x=1}^{|\Omega_X|}\nu_x\right)}$$

où ρ est la fonction Gamma.

A présent, calculons les lois *a posteriori* des paramètres π, A, ϕ. Nous allons voir que les lois *a posteriori* restent des lois Dirichlet. Commençons par la loi *a posteriori* de π.

2.3.3.7 Loi initiale

On note π comme la loi initiale de la chaîne.

$$\begin{aligned}
\mathbb{P}(\pi = p|Y^N = y^N, X^N = x^N, A = a, \phi = \chi) &= \mathbb{P}(\pi = p|X_0 = x_0) \\
&\propto \mathbb{P}(X_0 = x_0|\pi = p)\mathbb{P}(\pi = p)
\end{aligned}$$

On sait que $\mathbb{P}(X_0 = x_0|\pi = p)$ suit une loi Multinomiale de paramètres $(1, p)$. Cette loi Multinomiale n'a qu'un tirage car nous n'avons qu'une seule observation de la chaîne. Donc

$$\begin{aligned}\mathbb{P}(X_0 = x_0|\pi = p)\mathbb{P}(\pi = p) &\propto p(x_0) \times (|\Omega_X|)! \\ &\propto p(x_0) \times (|\Omega_X|)! \times \prod_{x=1}^{|\Omega_X|} p(x)^{1-1} \\ &\propto \frac{p(x_0) \times \prod_{x=1}^{|\Omega_X|} p(x)^{1-1}}{B(1 + \mathbb{1}_{\{x_0=1\}}, ..., 1 + \mathbb{1}_{\{x_0=|\Omega_X|\}})}\end{aligned}$$

En effet, pour toute valeur prise par x_0, le vecteur $(1 + \mathbb{1}_{\{x_0=1\}}, ..., 1 + \mathbb{1}_{\{x_0=|\Omega_X|\}})$ a toutes ses coordonnées égales à 1, sauf une égale à 2. De plus, la fonction B est symétrique. Donc la valeur de $B(1 + \mathbb{1}_{\{x_0=1\}}, ..., 1 + \mathbb{1}_{\{x_0=|\Omega_X|\}})$ ne dépend pas de x_0.

Ainsi, à une constante de normalisation près, $\mathbb{P}(X_0 = x_0|\pi = p)\mathbb{P}(\pi = p)$ suit une loi de Dirichlet de paramètres $(1 + \mathbb{1}_{\{x_0=1\}}, ..., 1 + \mathbb{1}_{\{x_0=|\Omega_X|\}})$. Par conséquent, le paramètre π suit *a posteriori* la même loi de Dirichlet.

2.3.3.8 Loi de transition

Passons, maintenant au calcul de la loi *a posteriori* de la loi de transition A.

$$\begin{aligned}\mathbb{P}(A = a|Y^N = y^N, X^N = x^N, \pi = p, \phi = \chi) =& \mathbb{P}(A = a|X^N = x^N) \\ =& \frac{\mathbb{P}(X^N = x^N|A = a)\mathbb{P}(A = a)}{\mathbb{P}(X^N = x^N)}\end{aligned}$$

Puisque $\mathbb{P}(X^N = x^N|A = a) = \mathbb{P}(X_0 = x_0) \prod_{n=1}^{N} a(x_{n-1}, x_n)$, on a

$$\mathbb{P}(X^N = x^N|A = a)\mathbb{P}(A = a) \propto \mathbb{P}(X_0 = x_0) \prod_{x_{n-1}=1}^{|\Omega_X|} \prod_{x_n=1}^{|\Omega_X|} [a(x, x')]^{n_{x,x'}}$$

avec $n_{x,x'} = \sum_{n=1}^{N} \mathbb{1}_{\{x_{n-1}=x, x_n=x'\}}$ le nombre de transitions de l'état j à l'état i. Ainsi, on remarque que sachant X^N, les lignes de la matrice A restent indépendantes car il est possible de décomposer la probabilité *a posteriori* de la matrice A en un produit de probabilités sur les lignes de la matrice A.

$$\mathbb{P}(A(x, 1) = a(x, 1), ..., A(x, |\Omega_X|) = a(x, |\Omega_X|)|X^N = x^N, \pi(x_0) = \pi_{x_0}) \propto \prod_{x'=1}^{|\Omega_X|} [a(x, x')]^{n_{x,x'}}$$

Par conséquent, les lignes de la matrice de transition A suivent une loi *a posteriori* de Dirichlet de paramètres $(n_{x,1} + 1, ..., n_{x,|\Omega_X|} + 1)$.

2.3.3.9 Loi d'émission

Passons maintenant au calcul de la loi *a posteriori* de la loi d'émission ϕ.

$$\begin{aligned}\mathbb{P}(\phi=\chi|X^N=x^N,Y^N=y^N,A=a,\pi=p) =&\mathbb{P}(\phi=\chi|X^N=x^N,Y^N=y^N)\\ \propto&\mathbb{P}(Y^N=y^N|X^N=x^N,\phi=\chi)\mathbb{P}(\phi=\chi|X^N=x^N)\\ \propto&\mathbb{P}(Y^N=y^N|X^N=x^N,\phi=\chi)\mathbb{P}(\phi=\chi)\end{aligned}$$

Puisque $\mathbb{P}(Y^N=y^N|X^N=x^N,\phi=\chi)=\prod_{n=0}^{N}\chi(x_n,y_n)$, on a

$$\mathbb{P}(Y^N=y^N|X^N=x^N,\phi=\chi)\mathbb{P}(\phi=\chi)\propto\prod_{x=1}^{|\Omega_X|}\prod_{y\in\Omega_Y}[\chi(x,y)]^{m_{x,y}}$$

avec $m_{x,y}=\sum_{n=0}^{N}\mathbb{1}_{\{x_n=x,y_n=y\}}$ le nombre d'émissions de l'état x à l'état y. Ainsi, on remarque que les lignes de la matrice d'émission sont indépendantes car la loi *a posteriori* de ϕ peut s'écrire comme le produit de probabilités de chaque ligne. On obtient donc

$$\mathbb{P}(\phi(x,1)=\chi(x,1),...,\phi(x,y)=\chi(x,y)|X^N=x^N,Y^N=y^N)\propto\prod_{y\in\Omega_Y}[\chi(x,y)]^{m_{x,y}}$$

Par conséquent, les lignes de la matrice d'émission ϕ suivent une loi *a posteriori* de Dirichlet de paramètres $(m_{x,1}+1,...,m_{x,|\Omega_Y|}+1)$.

Par la suite il est possible de calculer la probabilité du premier état de la chaîne sachant les paramètres et les observations.

$$\begin{aligned}\mathbb{P}(X_0=x_0|Y^N=y^N,\pi=p,A=a,\phi=\chi) =&\rho_0(x_0)\\ =&\frac{\mathbb{P}(Y^N=y^N|X_0=x_0,\pi=p,A=a,\phi=\chi)\mathbb{P}(X_0=x_0|\pi=p)}{\mathbb{P}(Y^N=y^N|\pi=p,A=a,\phi=\chi)}\\ =&\mathbb{P}(Y_0=y_0|X_0=x_0,\lambda=\lambda_{\Omega_X})\\ &\times\frac{\mathbb{P}(Y_1=y_1,...,Y_N=y_n|X_0=x_0,\lambda=\lambda_{\Omega_X})\mathbb{P}(X_0=x_0|\pi=p)}{\mathbb{P}(Y^N=y^N|\pi=p,A=a,\phi=\chi)}\\ \propto&\chi(x_0,y_0)p_{x_0}\mathbb{P}(Y_1=y_1,...,Y_N=y_n|X_0=x_0,\lambda=\lambda_{\Omega_X})\\ \propto&\chi(x_0,y_0)p_{x_0}\beta_0(x_0)\end{aligned}$$

Enfin, voici le calcul de loi de transition de la chaîne sachant les observations et les paramètres.

$$\begin{aligned}\mathbb{P}(X_n=x_n|X_{n-1}=x_{n-1},Y^N=y^N,\lambda_{\Omega_X}) =&\frac{\mathbb{P}(Y^N=y^N|X_n=x_n,X_{n-1}=x_{n-1},\lambda_{\Omega_X})}{\mathbb{P}(Y^N=y^N|X_{n-1}=x_{n-1},\lambda_{\Omega_X})}\\ &\times\mathbb{P}(X_n=x_n|X_{n-1}=x_{n-1},A=a)\\ =&\frac{\mathbb{P}(Y^{n-1}=y^{n-1}|X_{n-1}=x_{n-1},\lambda_{\Omega_X})}{\mathbb{P}(Y^{n-1}=y^{n-1}|X_{n-1}=x_{n-1},\lambda_{\Omega_X})}\\ &\times\frac{\phi(x_n,y_n)\beta_n(x_n)a(x_{n-1},x_n)}{\mathbb{P}(Y_n=y_n,...,Y_N=y_N|X_{n-1}=x_{n-1},\lambda_{\Omega_X})}\\ =&\frac{\chi(x_n,y_n)\beta_n(x_n)a(x_{n-1},x_n)}{\mathbb{P}(Y_n=y_n,...,Y_N=y_n|X_{n-1}=x_{n-1},\lambda_{\Omega_X})}\\ \propto&\chi(x_n,y_n)\beta_n(x_n)a(x_{n-1},x_n)\end{aligned}$$

2.3.3.10 Pseudo code de l'échantillonneur de Gibbs

L'algorithme de Gibbs alterne entre la mise à jour des paramètres sachant la chaîne de Markov et la mise à jour de la chaîne sachant les paramètres.

Pour débuter l'algorithme, il faut initialiser les valeurs de la chaîne de Markov. Voici la $it^{\text{ième}}$ itération de l'algorithme :

1. On simule $\phi_{it}|...$ par ligne
   $\forall\, x \in \Omega_X,\ (\phi_{it}(x,1),...,\phi_{it}(x,\Omega_Y))|... \sim D(m_{it-1,x,1}+1,...,m_{it-1,x,|\Omega_Y|}+1)$.
2. On simule $A_{it}|...$ par ligne
   $\forall\, x \in \Omega_X,\ (A_{it}(x,1),...,A_{it}(x,|\Omega_X|))|... \sim D(n_{it-1,x,1}+1,...,n_{it-1,x,|\Omega_X|}+1)$.
3. On simule $\pi_{it}|... \sim D(1+\mathbb{1}_{\{x_{it-1,0}=1\}},...,1+\mathbb{1}_{\{x_{it-1,0}=|\Omega_X|\}})$
4. On simule $X_{it,0}|... \sim \phi_{it}(x_0,y_0)\pi_{it,x_0}\beta_0(x_0)$
5. On simule dans l'ordre croissant $\forall\, n \in \{1,...,N\},\ X_{it,n}|... \sim \phi_{it}(x_n,y_n)\beta_n(x_n)A_{it}(x_{n-1},x_n)$

D'après l'algorithme de Gibbs, $(X_{it,0}|...,X_{it,n}|...,\pi_{it}|...,A_{it}|...,\phi_{it}|...)$ converge vers la loi jointe. $(X_{\infty,0},X_{\infty,n},\pi_\infty,A_\infty,\phi_\infty)$.

La dernière itération de l'algorithme sert pour estimer les loi des paramètres de la chaîne de Markov. Pour estimer les paramètres de la loi de Dirichlet du paramètre π, il suffit de calculer les moyennes et les variances empiriques puis de résoudre les systèmes d'équations d'ordre 2. Il est possible d'estimer de façon analogue les paramètres des lois des lignes de ϕ et des lignes de A.

De façon similaire à ce qui a été fait avec l'algorithme EM, il est possible de mettre en oeuvre l'échantillonneur de Gibbs avec plusieurs chaînes de Markov cachées suivant la même loi pour une meilleure estimation des paramètres. Pour ajuster l'algorithme dans ce cadre, il faut remplacer l'étape 3 par

3 bis. On simule $\pi_{it}|... \sim D(1+\sum_{c=1}^{C}\mathbb{1}_{\{x_{c,0}=1\}},...,1+\sum_{c=1}^{C}\mathbb{1}_{\{x_{c,0}=|\Omega_X|\}})$.

et on répète les étapes 4 et 5 autant de fois qu'il y a de chaînes.

2.3.4 Discussion

L'algorithme EM et l'échantillonneur de Gibbs permettent d'estimer les paramètres des lois d'une chaîne de Markov cachée avec des approches différentes. L'algorithme de Gibbs estime la loi des paramètres, vus comme variables aléatoires, alors que l'algorithme EM estime les valeurs des paramètres les plus probables, c'est-à-dire ceux qui maximisent la vraisemblance. Le choix des lois *a priori* est crucial pour l'estimation via l'échantillonneur de Gibbs. Il présente l'avantage de décomposer un problème de dimension supérieure en plusieurs sous-problèmes et de déterminer des intervalles de confiance des estimateurs. Les deux méthodes d'estimation on été testées sur des chaînes de Markov cachées simulées pour estimer la valeur des paramètres. Elles aboutissent globalement au même résultat mais l'algorithme EM est plus rapide (Rydén et al., 2008) et plus précis. Quelques résultats sont détaillés dans l'annexe A. Les paramètres dans le cadre MHMM-DF seront estimés à l'aide de l'algorithme EM afin d'optimiser le temps de calcul par rapport à l'échantillonneur de Gibbs.

Chapitre 3

MHMM-DF pour la dynamique d'espèces avec stade caché

Dans ce chapitre, nous présentons les démarches qui m'ont conduit à proposer une structure de graphe de dépendance pour représenter la dynamique spatio-temporelle des adventices. Afin de modéliser la dynamique locale, il faut prendre en compte le processus de survie de la banque de graines, le processus de germination, le processus de survie de la flore levée ainsi que le processus de production des nouvelles graines qui ne sont pas dispersées.

Une fois la dynamique locale des adventices modélisée, il reste à considérer la dynamique régionale, c'est-à-dire la colonisation entre champs dans le cas des adventices. Le modèle créé pour tenir compte de ces dépendances supplémentaires, que nous avons nommé Multi Hidden Markov Model with Data-Feedback (MHMM-DF), est défini par son graphe et les lois de transition du graphe. Le graphe du MHMM-DF est composé de noeuds observables et non-observables, représentant respectivement les populations observables et non-observables, ainsi que d'arêtes entre les noeuds correspondant aux processus. J'illustre dans la suite comment les noeuds et les arrêtes de ce graphe peuvent s'associer à la dynamique locale et régionale d'espèces avec stade non-observé à travers quelques exemples.

Si ce modèle a été élaboré dans le but de modéliser la dynamique des adventices, il est possible de l'appliquer à de nombreuse espèces avec stade non-observable. La dynamique des adventices étant moins complexe que la dynamique d'une espèce quelconque avec stade caché, nous avons ajouté des dépendances supplémentaires au modèle pour intégrer, par exemple, les processus de migration et de survie de la population observable. Le modèle de la dynamique des adventices est donc un sous modèle du MHMM-DF complet.

3.1 Modélisation de la dynamique locale des espèces avec stade caché

3.1.1 Limites du cadre HMM classique

On utilisera les notations définies dans le chapitre 9 à la p140.

Pour l'instant, on considère que les populations d'une espèce d'un patch sont indépendantes des autres patchs, pour mieux appréhender la structure du modèle associé à la dynamique locale de l'espèce. La modélisation de celle-ci doit avoir une structure de dépendance qui traduit toutes les interactions possibles entre les deux types de populations locales de l'espèce. Un processus biologique qui influence l'état d'une population dans un patch doit être pris en compte par la présence d'une arête spécifique dans le graphe.

Considérons un HMM où la suite de variables observées est entièrement définie par la suite de variables cachées (figure(3.1)), en considérant que les populations non-observables et les populations observables correspondent respectivement aux variables cachées et observées. Cette structure de dépendances ne peut s'appliquer à la dynamique des espèces avec stade caché. En effet, la survie de la population cachée est bien modélisée à travers la matrice de transition de la chaîne de Markov (figure 3.1 flèches noires) et l'effet de la population cachée sur la population observée est bien modélisée à travers la matrice d'émission (figure 3.1 flèches bleues). En revanche, l'état de la population cachée est indépendante de l'état de la population observée. La survie des populations observées n'est pas, elle aussi, modélisée.

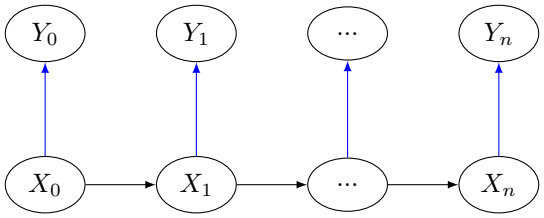


FIGURE 3.1 – Chaîne de Markov cachée

3.1.2 Introduction du data feedback et de la survie de la population observée

Borgy et al. (2015) ont étendu les chaînes de Markov cachées pour inclure l'influence de la population observée sur la population cachée du même pas de temps (figure 3.2 flèches rouges). De plus Borgy et al. (2015) on inclue un décalage de l'influence de la population caché vers la population observée. Ce modèle est adapté pour étudier la dynamique annuelle et locale des adventices car la population observée ne survit pas d'une année sur l'autre. On appellera ce modèle Hidden Markov model with data feedback (HMM-DF).

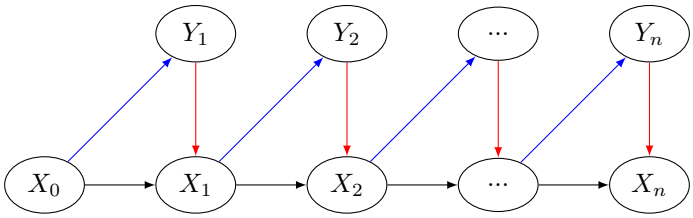


FIGURE 3.2 – Modèle de chaîne de Markov cachée avec retour des données associé à la dynamique locale des plantes annuelles

Seules les dynamiques avec renouvellement de la population observée à chaque pas de temps peuvent être modélisées par cette structure.

On peut modifier le graphe précédent pour considérer la survie des populations observables en ajoutant des arêtes entre les noeuds observables (figure 3.3). Cette structure de dépendance permet d'étudier la dynamique locale d'une plus grande variété d'espèces.

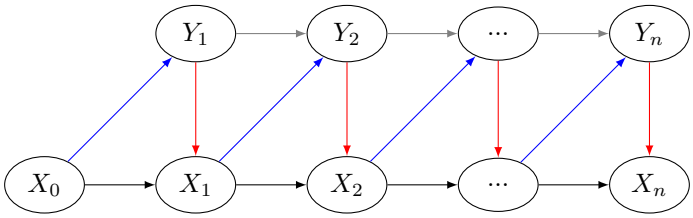


FIGURE 3.3 – Modèle de chaîne de Markov cachée avec retour des données et survie des populations observées.

3.2 Graphe de la dynamique spatiale d'espèces avec stade caché

Les graphes de dépendance précédents ne permettent pas de représenter les dépendances entre patchs. La dépendance entre patchs peut intervenir de 4 manières différentes, à savoir entre populations cachées et populations observables, dans un sens ou dans l'autre. Plusieurs extensions au HMM ont déjà été élaborées pour inclure une dépendance entre patchs, par exemple, les Factorial HMM (Ghahramani and Jordan (1997)) (figure 3.4) ou les Coupled HMM (Brand et al. (1997)) (figure 3.5). Le FHMM permet une colonisation de populations cachées vers populations observées alors que le CHMM permet une colonisation de populations cachées vers populations cachées.

Les figures ci-dessous ne prennent en compte que 2 patchs pour une meilleure lisibilité.

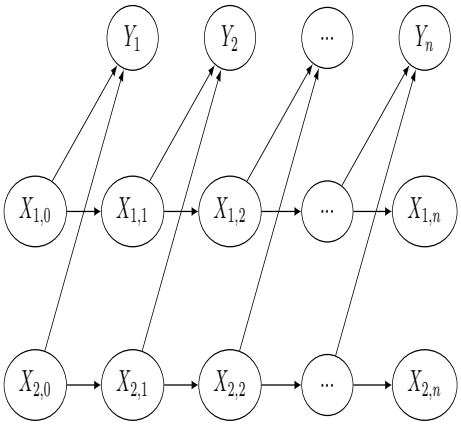


FIGURE 3.4 – Factorial Hidden Markov Model

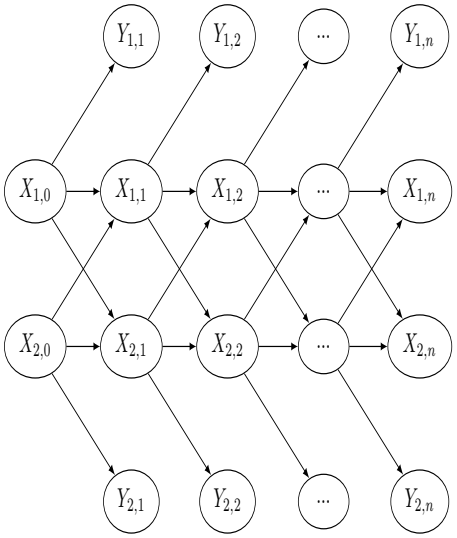


FIGURE 3.5 – Coupled Hidden Markov Model

Le problème majeur de ces modèles réside dans leur estimation par l'algorithme EM, d'une complexité algorithmique exponentielle en fonction du nombre de patchs. Si le nombre de patchs est trop grand, l'estimation ne peut être réalisée que de manière approchée Beal. (2003).

Un modèle avec une colonisation qui part uniquement des populations observées assure l'indépendance des chaînes cachées sachant les données, ce qui facilite l'estimation. Nous allons nous concentrer sur ce cas, qui correspond notamment à la dynamique des adventices.

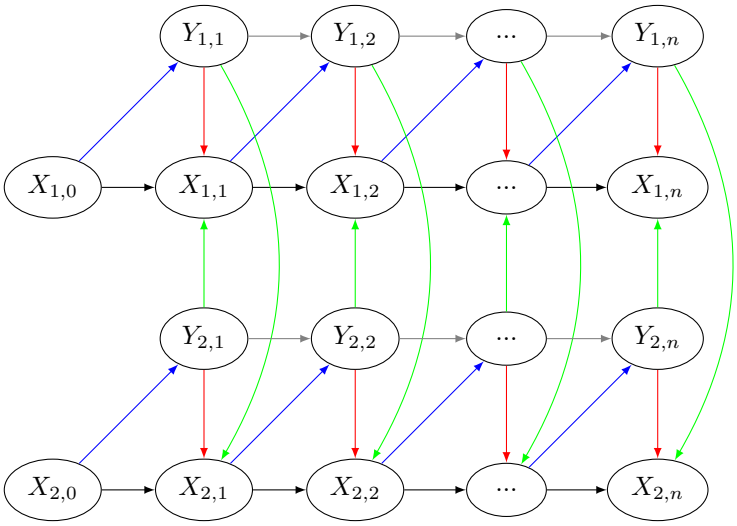


FIGURE 3.6 – Le graphe associé à la dynamique locale et régionale d'espèces avec stade caché qui comprend les processus de colonisation de populations observables vers populations cachées et de survie de la population observable.

Un modèle avec une colonisation qui part des populations observées vers les populations cachées (flèches vertes) tient compte du mode de colonisation des plantes, à savoir la production de graines par les plantes adultes vers la banque de graines.

Afin d'étudier la dynamique locale et régionale d'une plus grande variété d'espèces avec stade caché, nous avons étendu le modèle de la figure 3.6 à un modèle spatial qui inclut le processus de migration de populations observables vers les populations observables d'autres patchs (figure 3.7 flèches violettes).

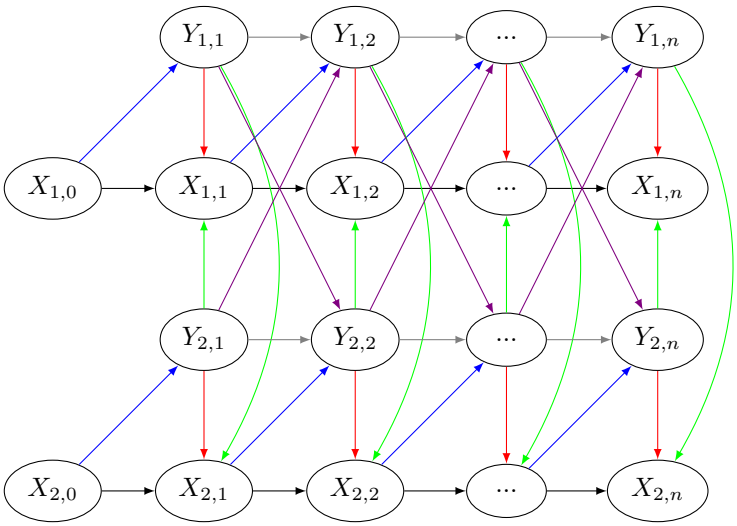


FIGURE 3.7 – Graphe associé a la dynamique locale et régionale des espèces avec stade caché avec colonisation partant de populations observables et survie des populations. Ce graphe correspond à la structure du MHMM-DF.

Le graphe de la figure 3.7 correspond à la structure de dépendance du MHMM-DF et permet

d'étudier la dynamique locale et régionale des espèces avec stade caché où seules les populations observées sont à l'origine de la colonisation et avec survie des deux populations d'un pas de temps à l'autre. Dans ce cadre, les populations cachées ne peuvent influencer directement une population d'un autre patch. Les espèces avec stade caché immobile répondent à ce type de dynamique.

3.3 Exemples d'application

Le graphe d'un MHMM-DF permet de représenter la dynamique d'une espèce avec deux stades, un caché et un observable. Ce modèle, ou un sous-modèle de celui-ci, peut être appliqué à beaucoup d'espèces. On donne ici plusieurs exemples d'espèces dont la dynamique peut être modélisée avec un MHMM-DF.

3.3.1 Dynamique des plantes

On considère que les plantes ont deux stades de vie : sous la forme d'une graine ou d'une plante adulte. Les plantes produisent des graines qui rentrent dans la banque de graines. Quand les conditions climatiques sont favorables et que la graine n'est plus dormante, la germination se produit. Une fois germée, la jeune plante va grandir jusqu'à devenir adulte. La figure 3.8 représente les interactions locales et régionales des plantes. La dynamique des plantes dans la figure 3.9 est modélisée à partir des populations de graines et les populations de plantes adultes. Les variables cachées $(X_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état de la population de graines sur la durée N dans les C patchs et les variables observées $(Y_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état de la population de plantes adultes dans les C patchs sur la durée N.

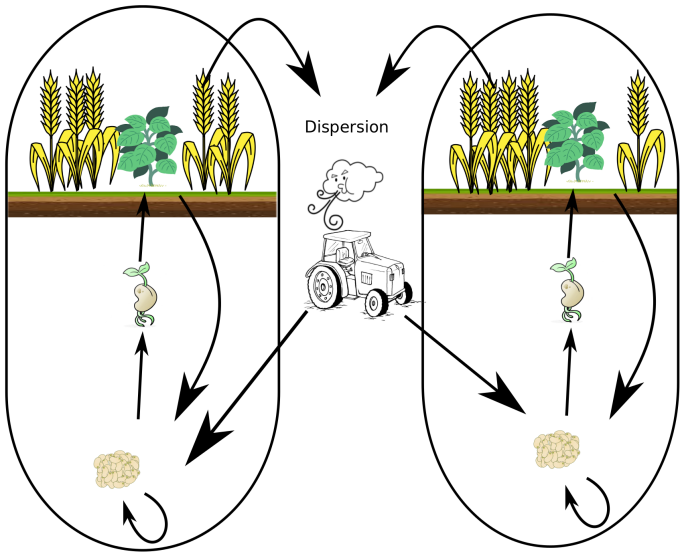


FIGURE 3.8 – Représentation des interactions locales et régionales entre population de graines et population de flore levée

Dans la figure 3.9, chaque flèche correspond à un processus dans la dynamique de la plante. Les flèches bleues correspondent au processus de germination et de survie jusqu'à l'âge adulte. Ainsi, cela dépend des ressources présentes dans le sol, de la culture semée ainsi que du type de désherbage utilisé. Les flèches vertes correspondent au processus de colonisation et les flèches rouges au processus de non-dispersion de nouvelles graines. Les facteurs faisant varier la colonisation sont

le vent, la taille et la surface de la graine ainsi que la distance entre les patchs. Les flèches noires correspondent au processus de survie de la population cachée et les flèches grises correspondent au processus de survie de la population observée.

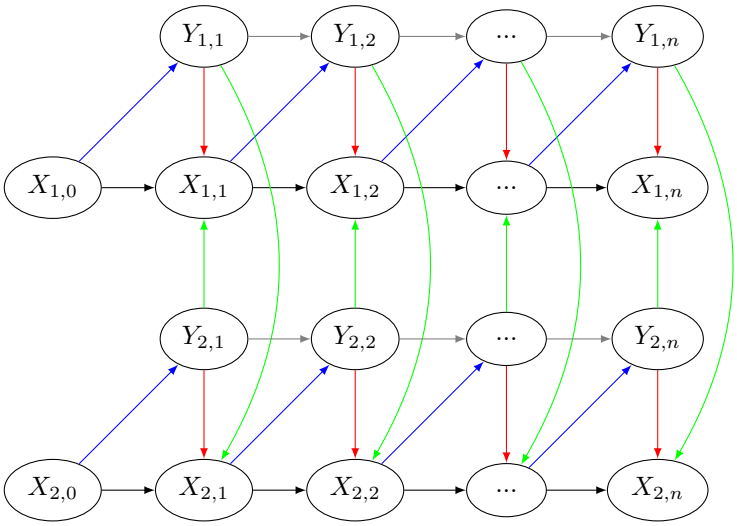


FIGURE 3.9 – Le sous-graphe d'un MHMM-DF associé à la dynamique des plantes

Dans ce modèle, on suppose que les flores levées peuvent survivre entre deux pas de temps. Cependant, si l'on s'intéresse aux adventices ou aux plantes annuelles et que l'on prend un pas de temps d'une année, alors la survie des plantes adultes n'a pas besoin d'être modélisée. Avec cette structure, on considère que les graines ne peuvent pas se déplacer d'un patch à un autre. Ainsi, on néglige l'effet des tracteurs qui peuvent transporter des graines d'adventices d'une banque de graines à une autre ou le transport de graines via une rivière.

3.3.2 Dynamique des puces

Le cycle de vie d'une puce est composé de 4 stades : l'oeuf, la larve, la pupe (le stade intermédiaire entre la larve et l'adulte) et la puce adulte. Les puces adultes pondent leurs oeufs sur la fourrure de l'hôte et ces oeufs peuvent tomber dans l'environnement local. Ils éclosent au bout de 2 à 5 jours pour former une larve. Le stade de larve dure entre 5 et 20 jours si les conditions sont favorables. Par la suite, les larves forment un cocon de soie. L'individu reste dans le cocon environ 1 à 4 semaines si les conditions sont adéquates. Si ce n'est pas le cas, le cocon peut rester dormant pendant une année. Une fois sorties du cocon, les puces ont besoin d'un hôte afin de se nourrir de son sang. L'éclosion du cocon dépend de l'humidité ambiante, de la température, des vibrations ainsi que du niveau de dioxyde de carbone. Une puce peut vivre jusqu'à 3 mois et la puce femelle peut pondre jusqu'à 5000 oeufs tout au long de sa vie. La dynamique des puces est représentée sur la figure 3.10.

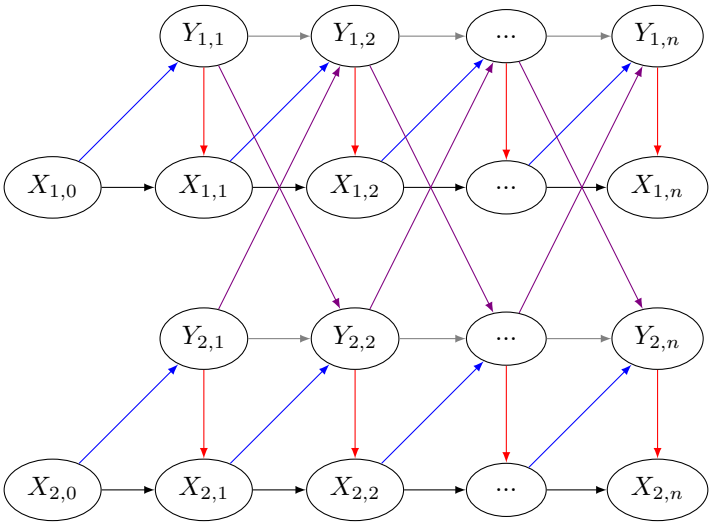


FIGURE 3.10 – Le sous-graphe d'un MHMM-DF associé associé à la dynamique des puces

Les variables cachées $(X_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état de la population d'individus cachés, ici sans hôte, au temps N dans le patch C et les variables observées $(Y_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état de la population visible, ici sur un hôte, d'individus dans le patch C au temps N.

Dans la figure 3.10, le processus d'émergence du cocon et d'infection d'un hôte est modélisé par les flèches bleues. Ainsi, la probabilité d'infection doit dépendre du nombre d'hôtes potentiels dans le patch c. Les flèches rouges représentent les oeufs qui tombent dans l'environnement local. La probabilité de survie de l'espèce sans hôte est modélisée avec les flèches noires entre variables cachées. La survie de l'espèce sur un hôte est modélisée par les flèches grises entre variables observées. L'interaction entre patchs se produit quand un hôte infecté change de patch. Cette interaction est modélisée par les flèches violettes. La figure 3.10 suppose que les individus sans hôte ne peuvent pas contaminer d'autres patchs. Cette hypothèse est cohérente à condition que les patchs soient assez éloignés.

3.3.3 Dynamique des escargots d'eau

L'espèce d'escargot d'eau Drepanotrema depressissimum peut persister dans l'environnement quand les conditions environnementales ne lui sont pas favorables (Lamy et al., 2013). L'espèce vit dans des zones arides où l'eau se fait rare. En période de sécheresse, les mares d'eau s'assèchent et les escargots sont forcés à estiver afin de survivre. Ils persistent en période de sécheresse en s'enfouissant dans le sol jusqu'aux prochaines pluies, ce qui les rend difficilement observables. Quand la mare contient de l'eau, les escargots peuvent se reproduire et coloniser d'autres mares si les mares sont reliées. Le graphe MHMM-DF de la dynamique des escargots d'eau est représenté sur la figure 3.11. Les escargots d'eau ont la particularité d'avoir une dynamique différente si la saison est favorable ou défavorable. En période de sécheresse tous les escargots estivent. Cependant, on suppose qu'en période favorable le nombre d'individus en forme adulte ou en forme d'estivation dépend de la quantité d'eau.

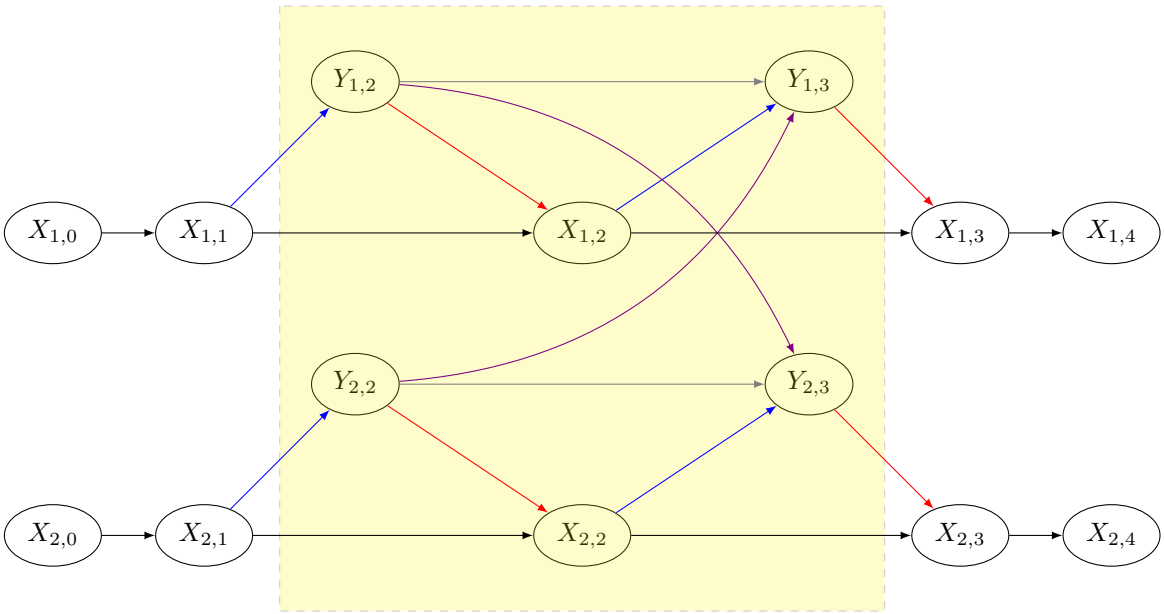


FIGURE 3.11 – Le graphe associé à la dynamique des escargots d'eau

Dans la figure 3.11, la zone jaune représente la saison des pluies. Les variables observées $(Y_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état des populations d'escargots adultes qui n'estivent pas. Les variables cachées $(X_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état des populations d'escargots qui estivent. Dans la figure 3.11, les flèches rouges correspondent au processus d'entrée en estivation. Les flèches bleues représentent la fin du processus d'estivation. Ce processus dépend de la quantité d'eau disponible dans les habitats. Les flèches violettes correspondent au processus de colonisation entre patchs. Le processus de colonisation dépend de l'existence ou non d'une connexion aquatique entre deux patchs. Les flèches noires représentent la survie des formes dormantes et les flèches grises représentent la survie des escargots qui n'estivent pas. Il est important de réaliser que les paramètres associés à la dynamique des escargots d'eau ne sont pas homogènes. En effet, les probabilités de transition dépendent de la quantité d'eau dans chaque patch. Prenons par exemple la probabilité de ne plus estiver. Si la prochaine période est une sécheresse, la probabilité de ne plus estiver sera 0. Cependant si la prochaine période n'est pas une sécheresse, cette probabilité sera strictement positive. On remarque que le graphe ne représente pas les observations en période de sécheresse, pendant lesquelles il n'y a que survie des populations dormantes. Pour représenter les observations en période de sécheresse, il suffit de leur affecter l'état d'extinction. De ce fait, le graphe de la dynamique des escargots d'eau est un sous-graphe du MHMM-DF.

3.3.4 Dynamique du parasite Myrmeconema neotropicum

Myrmeconema neotropicum (Dáttilo et al., 2013) est un nématode parasite qui utilise les fourmis Cephalotes atratus pour se reproduire. Les oiseaux mangent les fourmis infectées par le nématode et défèquent les oeufs du parasite. Les fourmis utilisent les excréments des oiseaux pour nourrir leurs larves. Les larves vont alors être infectées par le parasite qui va grandir dans la larve. Une fois que la larve se nymphose, le parasite va commencer à se reproduire dans l'abdomen. La transformation de la larve vers une fourmi adulte prend environ 2 à 3 mois. Une fois la transformation complétée, l'abdomen de la fourmi deviendra progressivement translucide puis rouge à cause des embryons du parasite. Plus l'abdomen est rouge vif, plus la fourmi infectée est âgée. Le parasite a la particularité de pouvoir contrôler partiellement son hôte. Ainsi, le parasite va forcer la fourmi à s'accrocher à

une feuille proche de baies rouges. Les baies rouges, qui sont une source de nourriture pour les oiseaux, ressemblent à l'abdomen des fourmis infectées. Ainsi les oiseaux vont manger les fourmis pensant que ce sont des baies rouges (Yanoviak et al., 2008). Le cycle du nématode peut ainsi continuer.

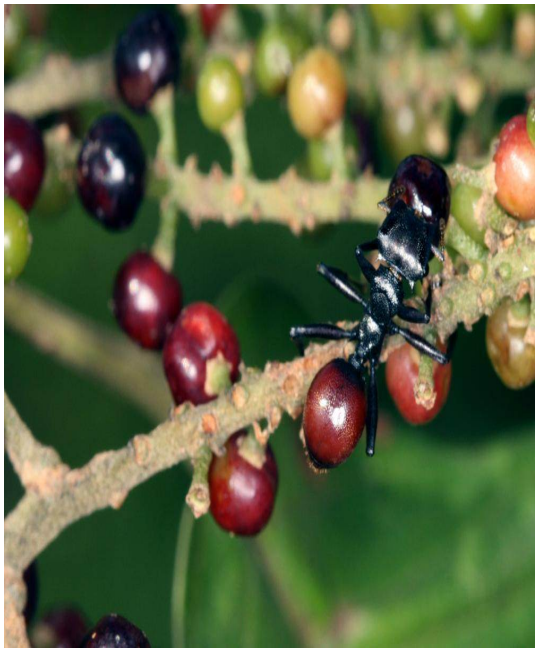


FIGURE 3.12 – Une fourmi infectée par le parasite Myrmeconema neotropicum au milieu de baies rouges.

La figure 3.12 montre une fourmi infectée par le parasite à côté de baies rouges. La dynamique locale et régionale du parasite Myrmeconema neotropicum est modélisée par MHMM-DF, dont le graphe est en figure 3.7.

Les variables observées $(Y_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état des fourmis adultes avec l'abdomen rouge au temps n dans le patch c et les variables cachées $(X_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état de l'infection auprès des larves et des fourmis avant qu'elles aient l'abdomen rouge au temps n dans le champ c. Chaque champ c correspond à une surface assez grande pour contenir l'intégralité du nid de colonie de fourmis. Les flèches rouges et vertes dans la figure 3.7 représentent le transport et l'infection d'une fourmi adulte infectée vers une ou des larves via un oiseau. Les flèches violettes représentent le processus de migration d'un adulte contaminé vers un autre patch. Le processus de migration est plausible car le nématode peut contrôler son hôte. Ce modèle suppose que la transmission de l'infection d'un patch à un autre à partir d'un hôte infecté caché n'est pas possible. Même si l'hôte venait à changer de patch afin de fourrager, la fourmi hôte retournera toujours à la colonie pour rapporter ses trouvailles. Les flèches bleues représentent la probabilité de croissance du nématode et de la larve vers une fourmi adulte avec un abdomen rouge. Les flèches noires représentent la survie des larves contaminées et les flèches grises représentent la survie des fourmis adultes infectées.

3.3.5 Dynamique des Ophiocordyceps

L'Ophiocordyceps est un champignon parasite des insectes (Andersen et al., 2012). Chaque espèce d'ophiocordyceps parasite un type d'espèce précis, comme des fourmis ou des chenilles. L'ophiocordyceps répand ses spores dans la forêt afin d'infecter de nouveaux hôtes. Une fois un hôte trouvé, les spores rentrent dans l'hôte par ses cavités et germent. Le temps de germination dépend de l'espèce d'ophiocordyceps et varie entre 1 jour et 1 mois. Pendant sa croissance, le champignon prend le contrôle sur les mouvements de son hôte. En prenant le contrôle de la fourmi, l'ophiocordyceps la force à grimper le long d'une plante. A environ 25 cm du sol, la fourmi y plante ses mandibules, ce qui restreint ses mouvements. L'emplacement choisi par le parasite est un endroit stratégique afin de compléter sa croissance et de disperser facilement ses spores une fois la croissance finie. Par la suite, la fourmi meurt lentement pendant que le champignon se nourrit d'elle afin de produire un ou plusieurs sporophores. Un sporophore est un appareil reproducteur qui permet la dispersion des spores.

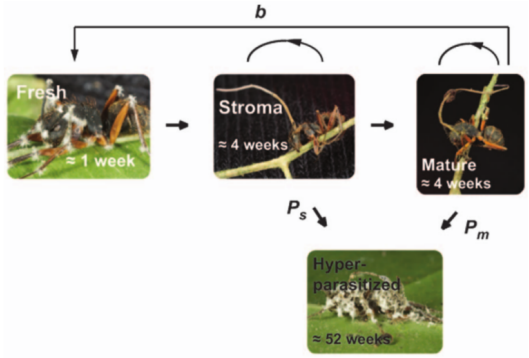


FIGURE 3.13 – Le cycle de vie des ophiocordyceps extrait de Andersen et al. (2012). b est la probabilité d'infecter de nouveaux hôtes, P_s est la probabilité que l'ophiocordyceps dans le stade stroma soit parasité et P_m est la probabilité que l'ophiocordyceps dans le stade mature soit parasité.

Le champignon reste ensuite dans son stade mature jusqu'à ce qu'il meure. La mort du champignon ophiocordyceps Unilateralis est souvent causée par un parasite champignon. Ce parasite champignon se développe en moyenne 4 semaines après maturité de l'ophiocordyceps unilateralis. Il stérilise l'ophiocordyceps Andersen et al. (2012). Le cycle de vie d'un champignon est représenté dans la figure 3.13

Afin d'appliquer le modèle à la dynamique locale et régionale des ophiocordyceps, on suppose que chaque patch représente une colonie de fourmis. Les variables cachées $(X_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état des individus infectés vivant dans le patch c au temps n. Ainsi l'état des résidus infectieux peut correspondre à une quantité de spores dans le patch c. Les variables observées $(Y_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent à l'état des individus morts infectés au temps n dans le patch c. La figure 4.1 représente le graphe associé à la dynamique des ophiocordyceps.

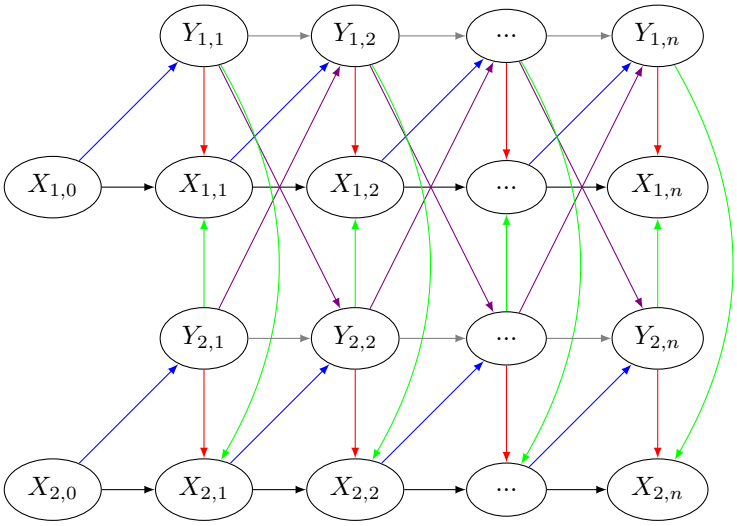


FIGURE 3.14 – Le graphe associé à la dynamique des Ophiocordyceps

Les flèches rouges de la figure 3.14 représentent la probabilité d'infection locale d'un nouvel hôte et les flèches vertes correspondent à la probabilité d'infection d'un hôte dans un patch voisin via colonisation. Il est très important de prendre en compte le nombre d'individus non infectés dans chaque patch dans la dynamique des ophiocordyceps. En effet, le nombre d'individus infectés va dépendre du nombre d'individus soumis au risque d'une infection. Ainsi, la probabilité d'infection doit dépendre du nombre d'individus non infectés. Les flèches bleues correspondent à la probabilité de développement du parasite en tuant son hôte. Les flèches noires représentent la survie des hôtes infectés par le champignon et les flèches grises représentent la survie du champignon dans un hôte

mort. Les champignons ne sont pas capables de se déplacer d'un patch à un autre. Cependant, quand un champignon parasite de fourmis est trop près d'une colonie de fourmis, les fourmis vont déplacer le champignon afin que l'infection de ne propage pas. Les flèches violettes représentent la probabilité qu'une colonie de fourmis déplace le champignon dans un cimetière. Ainsi cette probabilité doit dépendre de la distance entre le champignon et la colonie de fourmis.

Le modèle ne représente pas toutes les interactions possibles puisqu'une fourmi infectée ne meurt pas tout de suite et va se déplacer pour trouver un endroit idéal pour le développement du champignon. Lors de ce processus de migration de l'insecte, il se peut que l'insecte change de patch. Ainsi, la colonisation de variables cachées vers des variables observées est en théorie possible. De plus, le champignon est très contagieux ainsi une fourmi pourrait en contaminer une autre dans un autre patch sans que de nouveaux sporophores soient produits. Cependant, le modèle suppose que ces interactions sont négligeables.

3.3.6 Dynamique d'un jeu théorique

On suppose que tous les joueurs ont un objectif caché. De plus, le long de la partie, la progression de victoire de chaque joueur évolue sans que ses adversaires en soient avertis. Chaque tour est représenté par l'action de tous les joueurs simultanément. On suppose que C représente le nombre de joueurs. Les variables cachées $(X_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent au pourcentage de complétude de l'objectif de l'individu c au temps n. Les variables observées $(Y_{c,n})_{(c,n)\in\{1,\dots,C\}\times\{1,\dots,N\}}$ correspondent aux actions prises par le joueur c au temps n. Le sous-graphe du MHMM-DF de la figure 3.15 correspond à la dynamique du jeu.

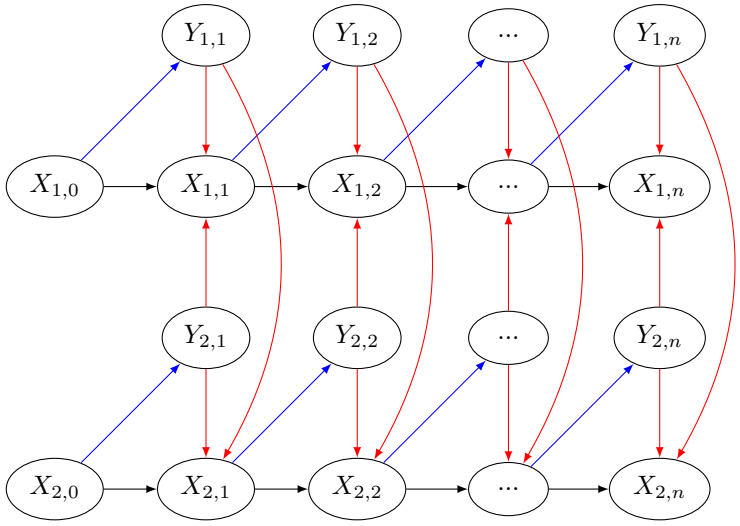


FIGURE 3.15 – Le graphe associé à un jeu théorique

Les flèches rouges dans la figure 3.15 représentent l'influence des actions prises par un joueur sur la progression de victoire de chacun des joueurs. Une flèche bleue correspond à une action entreprise par un joueur. Cette action reflète la stratégie du joueur afin d'atteindre le plus rapidement son objectif. Les flèches noires représentent la progression de victoire pour un joueur. Chaque pas de temps représente un tour. Le jeu s'arrête au tour N et le joueur étant le plus proche de ses conditions de victoire gagne.

Voici un exemple de jeu de type pierre papier ciseaux. On suppose qu'il n'y a que deux joueurs pour ce jeu. Les deux joueurs tirent à leur tour une carte sur laquelle sont affichés le nombre de victoires, le nombre de défaites et le nombre de matchs nuls que doit obtenir le joueur pendant la

partie. Les joueurs n'ont pas le droit de se montrer leur carte. La somme des victoires, défaites et match nuls sur chaque carte est égale à N. Après avoir tiré une carte, le jeu peut commencer. A chaque tour, les joueurs vont jouer à pierre papier ciseaux. A l'issue des N tours, le gagnant est le joueur avec le score le plus proche de ses conditions de victoires.

3.4 Discussion

Le graphe d'un MHMM-DF et ses sous-graphes permettent de modéliser la dynamique de plusieurs types d'espèces avec stade caché. Le MHMM-DF suppose que la population en stade caché ne peut pas influencer la population d'un patch voisin, contrairement aux CHMM et FHMM Ghahramani and Jordan (1997). Une telle hypothèse réduit la complexité algorithmique. Cependant, les espèces aquatiques, par exemple, qui pondent des oeufs dormants ne pourraient pas être modélisées car les courants des rivières peuvent déplacer les oeufs vers un autre patch. De plus, les espèces parasites, qui ne sont pas visibles une fois dans un hôte et ayant des hôtes très mobiles, ne peuvent être modélisées par un MHMM-DF. Pour les parasites, la population d'hôtes potentiels joue un rôle essentiel dans la dynamique du parasite. Si la population de l'hôte est connue à chaque pas de temps, il suffit de considérer que les observations sont des variables à deux dimensions qui incluent la population des hôtes potentiels.

Chapitre 4

Paramétrisation d'un MHMM-DF

Dans ce chapitre, nous allons voir plusieurs façons de modéliser les probabilités qui définissent un MHMM-DF afin d'étudier la dynamique d'une espèce avec stade non-observable. Pour cela, nous allons tout d'abord expliquer pourquoi paramétrer les lois du modèle puis nous allons exposer les avantages et les inconvénients d'utiliser des classes d'abondances. Par la suite, nous présenterons plusieurs options de modélisation. Enfin, nous établirons l'identifiabilité de certains des modèles paramétriques proposés.

4.1 Pourquoi a-t-on besoin de paramétrer les lois du MHMM-DF ?

Une modélisation non paramétrique correspond à une approche dans laquelle les probabilités du modèle ne reposent pas sur des familles de lois de probabilité paramétriques. Par conséquent, une probabilité non paramétrique du modèle comporte un paramètre pour chaque combinaison des états des variables impliquées dans cette probabilité de transition. Même si un modèle non paramétrique est souvent facile à implémenter, un grand nombre de données est nécessaire pour l'estimation. De plus, pour que tous les paramètres des lois de probabilité non paramétriques du modèle soient estimées, il faut que toutes les combinaisons des états des variables impliquées dans cette probabilité soient présentes dans les données. Quand peu de données sont à notre disposition, il est préférable de paramétrer les probabilités du modèle. Paramétrer les lois de transition permet d'avoir moins de paramètres à calculer, de réduire la variance des estimateurs et par conséquent d'avoir une meilleur capacité prédictive du modèle. Cependant, la modélisation d'une dépendance nécessite d'utiliser une loi appropriée. Si un processus biologique est modélisé par une loi inappropriée, les résultats risquent de ne pas être biologiquement cohérents et le modèle ne pourra pas être validé. Bien paramétrer les lois du modèle permet, une fois l'estimation faite, de comprendre la part relative de chaque processus dans la dynamique de l'espèce.

Il nous faut paramétrer 4 probabilités différentes, deux probabilités initiales π, ζ, une probabilité de transition A et une probabilité d'émission ϕ. Les probabilités (π, ζ, A, ϕ) caractérisent la loi des chaînes $(Y_{c,n})_{c\in\{1,\dots,C\},n\in\{1,\dots,N\}}$ et $(X_{c,n})_{c\in\{1,\dots,C\},n\in\{0,\dots,N\}}$. (π, ζ, A, ϕ) sont définies par

$$A(x_{c,n}, x_{c,n+1}, y_{n+1}^C) = \mathbb{P}(X_{c,n+1} = x_{c,n+1} | X_{c,n} = x_{c,n}, Y_{n+1}^C = y_{n+1}^C)$$

$$\phi(x_{c,n}, y_{c,n+1}, y_n^C) = \mathbb{P}(Y_{c,n+1} = y_{c,n+1} | X_{c,n} = x_{c,n}, Y_n^C = y_n^C)$$

$$\pi(x_{c,0}, y_0^C) = \mathbb{P}(X_{c,0} = x_{c,0} | Y_0^C = y_0^C)$$

$$\zeta(y_{c,0}) = \mathbb{P}(Y_{c,0} = y_{c,0})$$

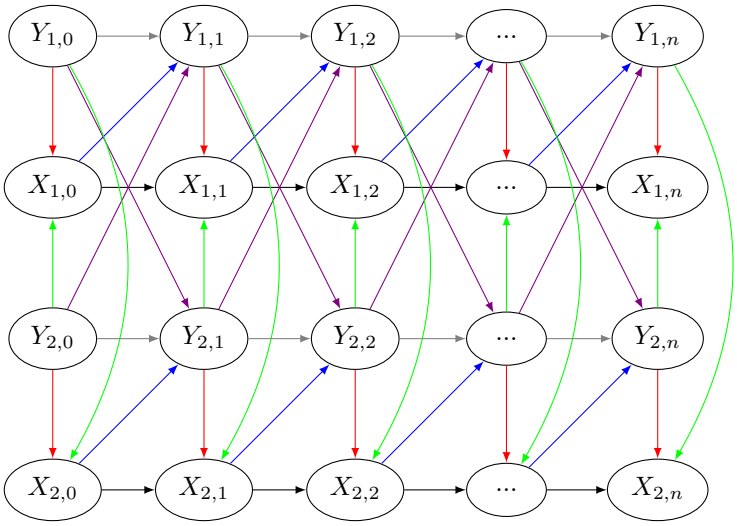


FIGURE 4.1 – Le graphe du MHMM-DF

4.2 Pourquoi des classes d'abondance ?

Les données binaires sont souvent utilisées et peuvent l'être dans notre cadre pour obtenir des prédictions sur la présence ou l'absence d'une population non observable dans un patch à un temps donné. Mais pour obtenir des prédictions plus précises sur l'état des variables cachées, il est important d'utiliser des données plus précises sur les variables observables que des données de type présence/absence. Une solution est l'utilisation de classes d'abondance, qui permet une représentation moins grossière qu'avec un modèle binaire tout en étant un compromis à l'utilisation directe des données chiffrées sur le nombre d'individus des populations étudiées. L'utilisation des données sur le nombre d'individus observables s'avère délicate. Tout d'abord, la récolte du nombre exact est difficile. De plus, la complexité de l'estimation à l'aide de l'algorithme EM est quadratique en le nombre de classes d'abondance de la population non observable.

Nous allons ainsi utiliser un nombre fixe de classes d'abondance pour les variables observées et cachées. On note $\Omega_X = \{0, ..., |\Omega_X| - 1\}$ l'ensemble des états possibles des variables cachées et $\Omega_Y = \{0, ..., |\Omega_Y| - 1\}$ celui des variables observables. On choisira les lois que suivront ces variables parmi des lois à support fini. Il en existe peu : la loi Binomiale, la loi Bêta-Binomiale, la loi Poisson Binomiale et la loi Géométrique. Le nombre de succès d'une série de tirages de Bernoulli indépendants de même probabilité de succès suit une loi Binomiale. La loi Bêta-Binomiale décrit le résultat d'une série de tirages de Bernoulli indépendants de probabilité de succès suivant une loi Bêta. La loi Poisson Binomiale correspond au processus de tirages de Bernoulli indépendants de probabilité de succès non constante. En plus du nombre d'états n, c'est-à-dire le nombre de tirages, la loi Binomiale repose sur 1 paramètre, la loi Béta-Binomiale dépend de 2 paramètres et la loi Poisson Binomiale utilise n autres paramètres.

4.3 Agrégation des observations

L'une des questions importantes dans le cadre MHMM-DF concerne la modélisation de l'influence entre chaînes. Peut-on distinguer l'influence de chaque chaîne ? Toutes les chaînes s'influencent-elles entre elles ou faut-il définir un voisinage de chaînes influençant une chaîne donnée ?

Distinguer l'influence de chaque chaîne reviendrait à lui attribuer un paramètre dans la probabilité de transition, ce qui ne serait pas raisonnable si le nombre de chaînes est trop grand. La

matrice de transition A de taille $|\Omega_X||\Omega_Y|^C \times |\Omega_X|$, grandit de façon exponentielle avec C et pour $C = 10$, $|\Omega_X| = |\Omega_Y| = 5$, A aurait 5^{11} lignes. Ainsi, il est préférable de ne pas distinguer l'influence de chaque chaîne.

On peut choisir d'agréger l'influence des chaînes sur une chaîne donnée. L'une des approches que nous avons essayé d'implémenter consiste à regrouper des zones d'influence en trois groupes, selon le poids de leur contribution à l'influence sur la chaîne considérée. Cette approche dans le cadre des adventices est détaillée dans l'annexe B, pour laquelle on considère la distance entre chaînes pour déterminer leur influence mutuelle. Dans un cadre plus général, il faudrait soit utiliser une fonction de dispersion qui necessite un paramètre suplémentaire soit définir une "distance" entre chaînes cohérente avec le système que l'on veut modéliser. Dans la suite de chapitre, on supposera que toutes les chaînes s'influencent entre elles de façon égale. Lors de l'analyse des données d'Epoisses au Chapitre 7, on déterminera un voisinage d'influence pour chaque patch plus approprié à l'étude des adventices.

4.3.1 Agrégation moyennée

Soit $c \in \mathcal{C}$. On pose $f_1 : \Omega_Y^{\mathcal{C}\setminus c} \to \Omega_Y$ la fonction d'agrégation moyennée définie par

$$\forall y_n^{\mathcal{C}\setminus c} \in \Omega_Y^{\mathcal{C}\setminus c},\ f_1(y_n^{\mathcal{C}\setminus c}) = \left\lceil \left(\frac{1}{C-1} \sum_{c' \in \mathcal{C}\setminus c} y_{c',n} \right) - 0.5 \right\rceil$$

Plusieurs combinaisons différentes d'états observés peuvent donner la même moyenne agrégée mais l'agrégation moyennée est rapide à exécuter et n'est pas coûteuse en complexité mémoire.

4.3.2 Agrégation alphabétique

L'agrégation alphabétique dépend du nombre de chaque état observé parmi les chaînes voisines. Par conséquent, toutes les chaînes sont considérées de la même manière.

Soit $g : \Omega_Y^{\mathcal{C}\setminus c} \to \Omega_Y^{\mathcal{C}}$ la fonction qui trie dans l'ordre décroissant un vecteur de $\Omega_Y^{\mathcal{C}\setminus c}$ et rajoute un zéro à la fin. Par exemple, quand $C = 4$ et $|\Omega_Y| = 5$ si les chaînes voisines ont comme états $(1, 4, 1)$ alors $g(Y^{\mathcal{C}\setminus c}) = g((1, 4, 1)) = (4, 1, 1, 0) = (g_1(Y^{\mathcal{C}\setminus c}), ..., g_C(Y^{\mathcal{C}\setminus c}))$.

Soit $f_2 : \Omega_Y^{\mathcal{C}\setminus c} \to \left\{0, ..., \binom{C+|\Omega_Y|-2}{C-1} - 1\right\}$ la fonction d'agrégation alphabétique définie par

$$\forall y_n^{\mathcal{C}\setminus c} \in \Omega_Y^{\mathcal{C}\setminus c},\ f_2(y_n^{\mathcal{C}\setminus c}) = \sum_{c'=C-1}^{1} \mathbb{1}_{\{g_{c'}(y_n^{\mathcal{C}\setminus c}) > g_{c'+1}(y_n^{\mathcal{C}\setminus c})\}} \sum_{j=g_{c'+1}(y_n^{\mathcal{C}\setminus c})+1}^{g_{c'}(y_n^{\mathcal{C}\setminus c})} \binom{|\Omega_Y| - j + c' - 1}{c' - 1}$$

Le coefficient binomial dans la somme correspond au nombre de $(c' - 1)$-combinaisons avec répétitions dans un ensemble à $|\Omega_Y| - j + 1$ éléments.

Voici un exemple. On suppose $Y^{\mathcal{C}\setminus c} = (4, 3, 3, 1)$ On a que

$$g(Y^{\mathcal{C}\setminus c}) = \begin{bmatrix} g_1(Y^{\mathcal{C}\setminus c}) = 4 \\ g_2(Y^{\mathcal{C}\setminus c}) = 3 \\ g_3(Y^{\mathcal{C}\setminus c}) = 3 \\ g_4(Y^{\mathcal{C}\setminus c}) = 1 \\ g_5(Y^{\mathcal{C}\setminus c}) = 0 \end{bmatrix}$$

On a que $f_2((4, 3, 3, 1)) = \binom{7}{3} + \binom{5}{2} + \binom{4}{2} + \binom{1}{0} = 35 + 10 + 6 + 1 = 53$

On peut montrer que f_2 est injective sur l'ensemble des listes triées d'éléments de Ω_Y. Ainsi, l'agrégation alphabétique permet de distinguer l'effet de deux combinaison d'états des chaînes

voisins qui ont la même moyenne mais pas des combinaisons avec le même nombre de chaînes dans chaque état. En contrepartie, ses complexités mémoire et algorithmique sont plus grandes. Aussi, la fonction f_2 est croissante pour l'ordre lexicographique. Par exemple, $f_2((5,1,1))$ est supérieur à $f_2((4,4,4))$. Cet ordre peut ne pas être représentatif de la réalité car il est possible qu'une configuration où tous les voisins prennent la valeur 4 ait plus d'influence que si l'un des voisins prend la valeur 5 et le reste prennent la valeur 1.

Dorénavant, nous utiliserons l'agrégation moyennée. Nous allons détailler comment modéliser les probabilités de transition du MHMM-DF.

4.4 Loi d'émission

La loi d'émission ϕ correspond à la probabilité de transition de la variable observée. Toutes les modélisations dans cette section sont détaillées dans le cadre MHMM-DF complet et par conséquent modélisent la loi de $(Y_{c,n+1}|X_{c,n}, Y_{c,n}, Y_n^{\mathcal{C}\setminus c})$. La modélisation de la loi d'émission pour le sous-graphe MHMM-DF pour la dynamique des plantes peut être extraite des modélisations présentées en supprimant l'influence entre variables observables. Dans ce cas, ϕ ne dépend pas des variables observées au temps précédent mais uniquement de la variable cachée et correspond à la probabilité de germination et de survie des plantes jusqu'à l'âge adulte.

4.4.1 Modélisation Binomiale

Pour $X_{c,n} = x_{c,n}$, $Y_{c,n} = y_{c,n}$ et $Y_n^{\mathcal{C}\setminus c} = y_n^{\mathcal{C}\setminus c}$ connues, on suppose que $Y_{c,n+1}$ suit une loi Binomiale de paramètres $(|\Omega_Y| - 1, p_{x,y^C})$. On a donc

$$\forall\, y \in \Omega_Y\ , \phi(x,y) = \binom{|\Omega_Y| - 1}{y} p_{x,y^C}^{y} (1 - p_{x,y^C})^{|\Omega_Y| - y - 1},$$

où la probabilité de succès de la loi Binomiale p_{x,y^C} dépend des états de $X_{c,n}, Y_{c,n}, Y_n^{\mathcal{C}\setminus c}$. Cette modélisation repose donc sur $|\Omega_X| \times |\Omega_Y|^C$ paramètres. Dans un contexte d'adventices, où la probabilité d'émission ne dépend que de l'état de la banque de graines, on a seulement $|\Omega_X|$ paramètres.

4.4.2 Modélisation Binomiale avec la fonction logistique

On notera cette modélisation (BL). La modélisation binomiale logistique permet d'avoir une moyenne croissante quand les variables explicatives grandissent. La modélisation Binomiale à l'aide de la fonction logistique suppose que $Y_{c,n+1}|X_{c,n}, Y_{c,n}, Y_n^{\mathcal{C}\setminus c}$ suit une loi Binomiale de paramètres $(|\Omega_Y| - 1, p_x)$ avec p_x défini par

$$p_x = \frac{1}{1 + \mathrm{e}^{-w_\mu}} \text{ avec } w_\mu = \mu_1 \frac{x}{|\Omega_X|} + \mu_2 \frac{y'}{|\Omega_Y|} + \mu_3 \frac{f(y'^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|} + \mu_0$$

où f est une fonction d'agrégation et $|f(\Omega_Y^{C-1})|$ est le nombre d'états possibles de la fonction d'agrégation.

L'avantage de la fonction logistique repose sur le fait qu'elle permet de n'avoir que 4 paramètres pour la loi d'émission. Ainsi, en simplifiant l'équation au-dessus on obtient

$$\mathbb{P}(Y_{c,n+1} = y | X_{c,n} = x, Y_{c,n} = y', Y_n^{\mathcal{C}\setminus c} = y'^{\mathcal{C}\setminus c}) = \binom{|\Omega_Y| - 1}{y} \left[\frac{1}{1 + \mathrm{e}^{-w_\mu}}\right]^{|\Omega_Y| - 1} \left[\mathrm{e}^{-w_\mu}\right]^{|\Omega_Y| - y - 1}.$$

Dans un contexte d'adventices avec un pas de temps d'une année, la modélisation repose seulement sur 2 paramètres et w_μ s'écrit $w_\mu = \mu_1 \frac{x}{|\Omega_X|} + \mu_0$.

Il est toujours possible d'inclure des covariables au sein de cette modélisation, par exemple pour modéliser l'effet de facteurs abiotiques sur l'abondance des adventices, en ajoutant un terme dans w_μ qui dépend d'un paramètre supplémentaire.

4.4.3 Modélisation Zéro inflated Binomiale avec la fonction logistique

On notera cette modélisation ($ZIBL$). Le Zéro inflated modèle, initialement utilisé par Lambert (1992), permet de modéliser des dynamiques où la population observable visite anormalement fréquemment un état donné, que l'on choisira être 0. On dira que la variable $Y_{c,n+1}|X_{c,n}, Y_{c,n}, Y_n^{\mathcal{C}\setminus c}$ suit une ZIB avec paramètres $(|\Omega_Y| - 1, p_x, p_{ZI})$ si pour tout $y \in \Omega_Y$,

$$\mathbb{P}(Y_{c,n+1} = y|X_{c,n} = x, Y_{c,n} = y', Y_n^{\mathcal{C}\setminus c} = {y'}^{\mathcal{C}\setminus c}) = \mathbb{1}_{\{y=0\}} p_{ZI} + (1-p_{ZI}) \binom{|\Omega_Y| - 1}{y} [p_x]^y [1-p_x]^{|\Omega_Y|-y-1}$$

avec $p_x = \frac{1}{1+\mathrm{e}^{-w_\mu}}$ et $w_\mu = \mu_1 \frac{x}{|\Omega_X|} + \mu_2 \frac{y'}{|\Omega_Y|} + \mu_3 \frac{f({y'}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|} + \mu_0$. Cette modélisation a 5 paramètres. Dans un contexte de plantes annuelles, les paramètres μ_2 et μ_3 ne sont pas présent.

4.4.4 Modélisation Binomiale uniquement pour les états cachés non éteints

On notera cette modélisation (BU). A l'aide d'une version modifiée du modèle Zéro inflated Binomiale, nous avons construit une probabilité de transition qui permet de ne pas produire d'observation quand l'état de la variable cachée est dans le premier état. Il nous suffit de fixé le paramètre de la Zéro inflated à $p_{ZI} = \mathbb{1}_{\{x=0\}}$. Cette modélisation est appropriée à l'étude des adventices puisque qu'elle permet d'imposer que si la banque de graines est dans l'état d'extinction alors aucune flore levée ne peut être observée. On peut remarquer que si l'on passe les données sur les variables cachées au logarithme dans une modélisation binomiale logistique, la modélisation serait très similaire à celle de la modélisation BU. En effet, si $X_{c,n} = 0$ alors, $log(X_{c,n}) = -\infty$, par conséquent, le paramètre de la binomiale logistique vaudrait 0 ce qui impliquerait que la variable observée prendrait obligatoirement la valeur 0.

4.5 Loi de transition

La loi de transition A correspond à la probabilité de transition de la variable cachée. Toutes les modélisations dans cette section modélisent la loi de $X_{c,n+1}|Y_{n+1}^{\mathcal{C}\setminus c}, X_{c,n}, Y_{c,n+1}$ pour le MHMM-DF complet. Plusieurs méthodes sont possibles afin de modéliser la loi de transition A. On peut choisir de modéliser séparément chaque processus ou de modéliser tous les processus directement à l'aide d'une distribution connue. Ces deux méthodes seront explorées.

Pour la dynamique des plantes, la loi de transition A correspond à la combinaison des processus de survie des graines au temps $n-1$, de colonisation et de production locale de graines non dispersées.

4.5.1 Modélisation en séparant les processus

4.5.1.1 Modélisation de la survie des graines séparément de l'approvisionnement de nouvelles graines

On notera la modélisation Binomiale de la survie des graines séparément de l'approvisionement des nouvelles comme $(BSSA)$. Elle permet de décomposer la loi de transition en deux processus. L'un des processus traduit l'influence des variables observées sur la variable cachée et l'autre processus dépend des variables cachées. Ils sont tous deux modélisés à l'aide de lois Binomiales. On pose $U_{1,n}$ et $U_{2,n}$ deux variables aléatoires qui suivent respectivement une loi Binomiale de paramètres $(X_{c,n} = x', p_y)$ et $(|\Omega_X| - 1, p_{y,y'})$ et on pose

$$X_{c,n+1} = max(U_{1,n}, U_{2,n}).$$

Le paramètre de la loi Binomiale p_y dépend de la variable cachée $X_{c,n}$ et de la variable observable $Y_{c,n+1}$ alors que $p_{y,y'}$ ne dépend que des variables observées Y^C_{n+1}. On a donc pour tout $x \in \Omega_X$,

$$\begin{aligned}
\mathbb{P}(X_{c,n+1} = x | Y^{\mathcal{C}\setminus c}_{n+1} = y', X_{c,n} = x', Y_{c,n+1} = y) = &\binom{x'}{x} p_y^x (1-p_y)^{x'-x} \\
&\times \sum_{z'<x} \binom{|\Omega_X|-1}{z'} p_{y,y'}^{z'} (1-p_{y,y'})^{|\Omega_X|-1-z'} \\
&+ \binom{|\Omega_X|-1}{x} p_{y,y'}^{x} (1-p_{y,y'})^{|\Omega_X|-1-x} \\
&\times \sum_{z'<x} \binom{x'}{z'} p_y^{z'} (1-p_y)^{x'-z'} \\
&+ \binom{x'}{x} p_y^{x} (1-p_y)^{x'-x} \\
&\times \binom{|\Omega_X|-1}{x} p_{y,y'}^{x} (1-p_{y,y'})^{|\Omega_X|-1-x}
\end{aligned}$$

Le premier terme correspond au cas $X_{c,n+1} = U_{1,n} = U_{2,n}$, le second au cas $X_{c,n+1} = U_{1,n}$ et $U_{1,n} > U_{2,n}$ et le dernier au cas $X_{c,n+1} = U_{2,n}$ et $U_{2,n} > U_{1,n}$. Afin d'incorporer le moins de paramètres possible dans cette probabilité de transition, on peut définir les paramètres des lois Binomiales comme des fonctions logistiques ayant un paramètre pour chaque dépendance. Dans le cadre de la dynamique des adventices, cette modélisation sépare la survie des graines de l'approvisionnement de nouvelles graines provenant du champ local et/ou de la colonisation.

4.5.1.2 Modélisation Sparse

Le modèle sparse se base sur une idée très simple. Il considère que la variation de l'état de la variable cachée n'est pas trop importante. Plus précisément, pour tout $n \in \{1, \dots, N\}$ et $c \in \mathcal{C}$, on a $|X_{c,n+1} - X_{c,n}| \leq 1$.

On suppose que tout état des variables observables est une classe d'abondance associée à un intervalle, dont on utilise les bornes afin de déterminer l'état de la variable cachée. Par conséquent, cette modélisation dépend du découpage des données en classes. Pour tout $c \in \mathcal{C}$, on pose $m_{y_{c,n+1}}$ et $M_{y_{c,n+1}}$ les bornes inférieures et supérieures de l'intervalle associé à l'état $y_{c,n+1}$. On pose aussi

$$m_{y^{\mathcal{C}\setminus c}_{n+1}} = \frac{\sum_{c' \in \mathcal{C}\setminus c} m_{y_{c',n+1}}}{C-1} \text{ et } M_{y^{\mathcal{C}\setminus c}_{n+1}} = \frac{\sum_{c' \in \mathcal{C}\setminus c} M_{y_{c',n+1}}}{C-1}.$$

On détermine l'intervalle d'influence possible provenant des variables observées comme $[B_-, B_+]$ avec

$$B_- = \nu_0 + \nu_2 m_{Y_{c,n}} + \nu_3 m_{y_{n+1}^{\mathcal{C}\setminus c}}$$
$$B_+ = \nu_0 + \nu_2 M_{Y_{c,n}} + \nu_3 M_{y_{n+1}^{\mathcal{C}\setminus c}}$$

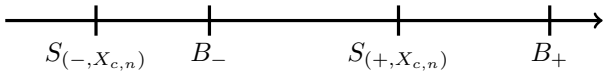
où ν_0, ν_2 et ν_3 sont des paramètres.

Cet intervalle est utilisé afin de déterminer, parmi trois états, l'état de la variable cachée au temps n selon l'influence de la variable cachée au temps $n-1$. Cette influence est définie à l'aide des seuils fixé au préalable de la loi log-normale que l'on notera $S_{(-,X_{c,n})}$ et $S_{(+,X_{c,n})}$. Par conséquent, la probabilité de transition a autant de loi log-normale que d'état de la variable cachée. Pour $x_{c,n}$ donnée, les paramètres de la loi log-normale sont $(\mu_{x_{c,n}}, \sigma)$. On suppose que σ est indépendant de la variable cachée.

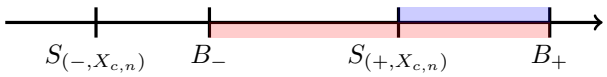
La probabilité de transition $\mathbb{P}(X_{c,n+1} = x_{c,n+1} | Y_{n+1}^{\mathcal{C}\setminus c} = y_{n+1}^{\mathcal{C}\setminus c}, Y_{c,n+1} = y_{c,n+1})$ est égale à

$$\begin{cases} 0 & \text{si } |x_{c,n+1} - x_{c,n}| > 1 \\ \dfrac{max(B_+, S_{(+,X_{c,n})}) - max(B_-, S_{(+,X_{c,n})})}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} + 1 \\ \dfrac{min(B_+, S_{(-,X_{c,n})}) - min(B_-, S_{(-,X_{c,n})})}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} - 1 \\ \dfrac{B_+ - S_{(-,X_{c,n})} + S_{(+,X_{c,n})} - B_- - \big(max(B_+, S_{(+,X_{c,n})}) - min(B_-, S_{(-,X_{c,n})})\big)}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} \end{cases}$$

Voici un exemple pour visualiser le calcul de la probabilité de transition.



Dans ce cas particulier, la probabilité que $X_{c,n+1} = x_{c,n} + 1$ peut se voir sur le dessin avec le rapport entre la partie bleue sur la partie rouge.



La modélisation sparse ne correspond pas à la dynamique des adventices à cause de la restriction des transitions des états possibles. Si on suppose que la variation de l'état de la variable cachée $|X_{c,n+1} - X_{c,n}|$ est inférieure ou égale à $\delta \geq 1$, on peut considérer 2δ seuils. Le cas à 4 seuils est détaillé en Annexe C.

4.5.2 Modélisation directe de tous les processus

4.5.2.1 Modélisation Binomiale logistique

On notera la modélisation Binomiale à l'aide de la fonction logistique (BL). Elle correspond à une loi Binomiale de paramètre de succès modélisé par une régression logistique dont les variables explicatives sont les états des variables qui influencent l'état de la variable d'arrivée. On suppose que $(X_{c,n+1} = x_{c,n+1} | , Y_{n+1}^{\mathcal{C}\setminus c} = y_{n+1}^{\mathcal{C}\setminus c}, X_{c,n} = x_{c,n}, Y_{c,n+1} = y_{c,n+1})$ suit une loi Binomiale de paramètres $(|\Omega_X| - 1, p_{(y_{n+1}^{\mathcal{C}\setminus c}, x_{c,n}, y_{c,n+1})})$ où la probabilité $p_{(y_{n+1}^{\mathcal{C}\setminus c}, x_{c,n}, y_{c,n+1})}$ est définie à l'aide de la fonction logistique :

$$p_{(y_{n+1}^{\mathcal{C}\setminus c}, x_{c,n}, y_{c,n+1})} = \frac{1}{1 + \mathrm{e}^{-w_\nu}}$$

avec $w_\nu = \nu_0 + \nu_1 \dfrac{x_{c,n}}{|\Omega_X|} + \nu_2 \dfrac{y_{c,n+1}}{|\Omega_Y|} + \dfrac{f(y_{n+1}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|} \nu_3$ et f est la fonction d'agrégation.

En simplifiant, on obtient donc

$$\mathbb{P}(X_{c,n+1}=x|Y_{n+1}^{\mathcal{C}\setminus c}=y_{n+1}^{\mathcal{C}\setminus c},X_{c,n}=x_{c,n},Y_{c,n+1}=y_{c,n+1})=\binom{|\Omega_X|-1}{x}\left[\frac{1}{1+\mathrm{e}^{-w_\nu}}\right]^{|\Omega_X|-1}\left[\mathrm{e}^{-w_\nu}\right]^{|\Omega_X|-x-1}$$

4.5.2.2 Modélisation Poisson Binomiale avec la fonction logistique

On notera (PBL) la modélisation Poisson Binomiale à l'aide de la fonction logistique. Une variable suivant une loi de Poisson Binomiale peut être vue comme la somme de plusieurs variables de loi de Bernouilli avec une probabilité de succès différente pour chaque variable. Cette méthode est importante car la probabilité de succès des variables de Bernouilli peut varier. Dans le contexte de la modélisation d'un processus de survie, il est important que l'état au temps suivant soit forcément plus petit ou égal à l'état précédent.

On suppose que $(X_{c,n+1}=x_{c,n+1}|,Y_{n+1}^{\mathcal{C}\setminus c}=y_{n+1}^{\mathcal{C}\setminus c},X_{c,n}=x_{c,n},Y_{c,n+1}=y_{c,n+1})$ suit une loi Poisson Binomiale $(|\Omega_X|-1,p_0,...,p_{(|\Omega_X|-1)})$. où pour tout $k\in\{0,...,|\Omega_X|-1\}$, p_k est la probabilité de succès de la k^e Bernouilli. La probabilité de transition A est

$$\mathbb{P}(X_{c,n+1}=x_{c,n+1}|,Y_{n+1}^{\mathcal{C}\setminus c}=y_{n+1}^{\mathcal{C}\setminus c},X_{c,n}=x_{c,n},Y_{c,n+1}=y_{c,n+1})=\sum_{A\in F_{x_{c,n+1}}}\prod_{i\in A}p_i\prod_{j\in\bar{A}}(1-p_j)$$

où $F_{x_{c,n+1}}$ est l'ensemble de tous les sous-ensembles de Ω_X contenant $x_{c,n+1}$ éléments et $\bar{A}$ est le complémentaire de A.

Dans le modèle ci-dessus, les $|\Omega_X|$ paramètres de succès de la loi Poisson Binomiale sont indépendants de $y_{n+1}^{\mathcal{C}}$ et de $x_{c,n}$. Dans ce cas, il y a $|\Omega_X|$ paramètres dans le MHMM-DF. Cependant, en ajoutant la dépendance aux variables qui influencent $X_{c,n+1}$, il y en aurait $|\Omega_X|^2|f(\Omega_Y^{C-1})||\Omega_Y|$. En supposant que chaque paramètre de succès est une fonction logistique, on peut réduire le nombre de paramètres à $4|\Omega_X|$ paramètres.

L'utilisation de la loi Poisson Binomiale permet de bien modéliser le processus de survie des variables cachées au sein de la probabilité de transition A pour la dynamique des adventices. Pour cela, on considère que la loi Poisson Binomiale n'a que 2 paramètres de succès où l'un des paramètres est dépendant de la survie des variables cachées et l'autre non. Cependant, les deux paramètres de succès partagent les mêmes paramètres au sein de la fonction logistique, sauf pour la survie des variables cachée, comme définit ci-dessous. Pour tout $k\in\{0,...,x_{c,n}\}$, on pose

$$p_k=\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_1\frac{x_{c,n}}{|\Omega_X|}+\nu_2\frac{y_{c,n+1}}{|\Omega_Y|}+\frac{f(y_{n+1}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|}\nu_3)}}$$

et pour tout $k\in\{x_{c,n}+1,...,|\Omega_X|-1\}$

$$p_k=\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_2\frac{Y_{c,n+1}}{|\Omega_Y|}+\frac{f(y_{n+1}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|}\nu_3)}}.$$

On remarque donc qu'il n'y a que 4 paramètres avec cette modélisation.

4.6 Identifiabilité générique de plusieurs modèles

Une chaîne de Markov cachée avec N pas de temps est dite génériquement identifiable si l'ensemble des choix de paramètres non identifiables est de mesure de Lebesgue nulle.

Tout d'abord, on suppose que les probabilités initiales sont paramétrées à l'aide de lois Binomiales avec la fonction logistique. Dans cette section, nous allons prouver l'identifiabilité du modèle selon le choix des probabilités ϕ et A.

Dans un premier temps, on suppose que

— la probabilité d'émission ϕ est paramétrée par une BL :

$$Y_{c,n+1}|X_{c,n} = x, Y_{c,n} = y', Y_n^{\mathcal{C}\setminus c} = {y'}^{\mathcal{C}\setminus c} \sim B(|\Omega_Y| - 1, p_x)$$

avec

$$p_x = \frac{1}{1 + \mathrm{e}^{w_\mu}} \text{ et } w_\mu = \mu_0 + \mu_1 \frac{x}{|\Omega_X|} + \mu_2 \frac{y'}{|\Omega_Y|} + \mu_3 \frac{f({y'}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|},$$

— la probabilité de transition A est paramétrée par une BL :

$$X_{c,n+1}|, Y_{n+1}^{\mathcal{C}\setminus c} = y_{n+1}^{\mathcal{C}\setminus c}, X_{c,n} = x_{c,n}, Y_{c,n+1} = y_{c,n+1} \sim B(|\Omega_X| - 1, p_{(y_{n+1}^{\mathcal{C}\setminus c}, x_{c,n}, y_{c,n+1})})$$

avec

$$p_{(y_{n+1}^{\mathcal{C}\setminus c}, x_{c,n}, y_{c,n+1})} = \frac{1}{1 + \mathrm{e}^{-w_\nu}} \text{ et } w_\nu = \nu_0 + \nu_1 \frac{x_{c,n}}{|\Omega_X|} + \nu_2 \frac{y_{c,n+1}}{|\Omega_Y|} + \nu_3 \frac{f(y_{n+1}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})$$

Ensuite, on prouve l'identifiabilité d'un modèle dans lequel

— la probabilité d'émission est paramétrée par une BU :

$$Y_{c,n+1}|X_{c,n} = x, Y_{c,n} = y', Y_n^{\mathcal{C}\setminus c} = {y'}^{\mathcal{C}\setminus c} \sim \begin{cases} 0 & \text{avec probabilité 1 si } X_{c,n} = 0 \\ B(|\Omega_Y| - 1, p_y) & \text{si } X_{c,n} \neq 0 \end{cases}$$

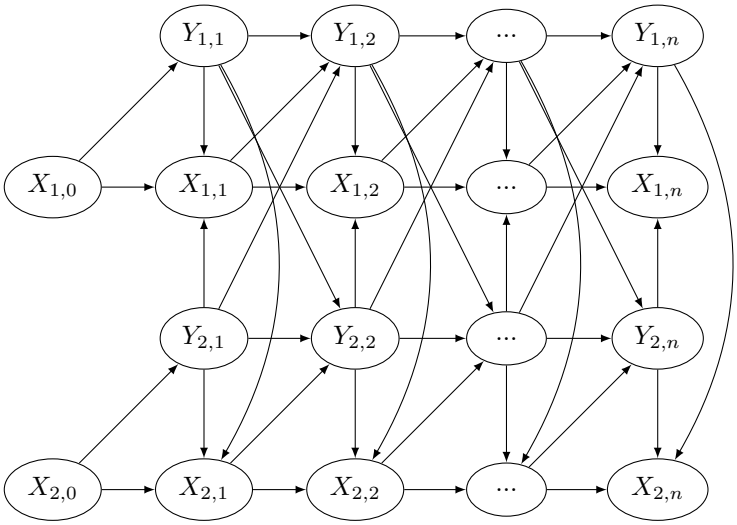
avec

$$p_x = \frac{1}{1 + \mathrm{e}^{w_\mu}} \text{ et } w_\mu = \mu_0 + \mu_1 \frac{x}{|\Omega_X|} + \mu_2 \frac{y'}{|\Omega_Y|} + \mu_3 \frac{f({y'}^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|},$$

— la probabilité de transition est paramétrée par une BL comme dans le premier cas.

Notons que la démonstration de l'identifiabilité générique lorsque la loi de transition A est une PBL est identique à celle pour une BL.

Montrons l'identifiabilité générique du MHMM-DF.



La preuve de l'identifiabilité générique repose sur le théorème 10 de Allman et al. (2009) :

Théorème 10 (Allman et al. (2009))**.** *Les paramètres d'une chaîne de Markov cachée avec r états cachés et s états observables sont génériquement identifiables à partir de la distribution marginale de $2L+1$ variables consécutives si*

$$\binom{L+s-1}{s-1} \geq r.$$

Ce résultat est valide pour une chaîne de Markov cachée. Il est possible de convertir le MHMM-DF en une chaîne de Markov cachée avec H_n les variables cachées et O_n les variables observées (voir Figure 4.2) en réunissant les variables observées et les variables cachées par pas de temps $H_n = (X_n^C, Y_{n+1}^C)^t$, et en groupant les variables observées $O_n = Y_{n+1}^C$. On remarque que cette conversion en chaîne de Markov cachée duplique les variables observées. On note ϕ_{hmm} et A_{hmm} respectivement la matrice d'émission et la matrice de transition de la chaîne de Markov cachée. ϕ_{hmm} est déterministe et A_{hmm} dépend des probabilités ϕ et A du MHMM-DF.

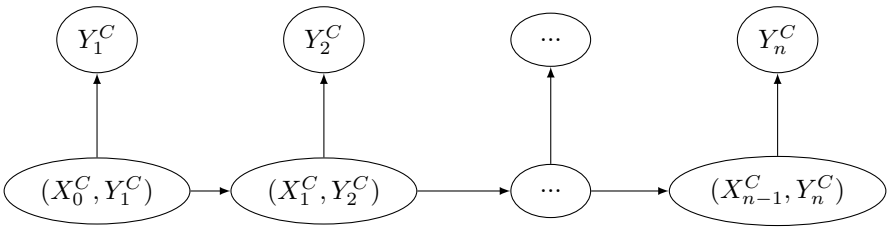


FIGURE 4.2 – Conversion d'un MHMM-DF en un HMM.

En appliquant le théorème à la chaîne de Markov cachée définie par (H_n, O_n), on montre la proposition suivante.

Propriété 11. *Les paramètres (ϕ_{hmm}, A_{hmm}) de la chaîne de Markov cachée correspondant à la conversion du MHMM-DF sont identifiables à partir de 7 variables observables consécutives si $|\Omega_X| \leq |\Omega_Y|$ et $C > 2$.*

Démonstration. Le nombre d'états cachés est $r = |\Omega_X|^C |\Omega_Y|^C$ et le nombre d'états observables est $s = |\Omega_Y|^C$. On utilise le théorème 6 de Allman et al. (2009) avec $L = 3$. Il suffit donc de vérifier que

$$\binom{|\Omega_Y|^C + 2}{|\Omega_Y|^C - 1} \geq |\Omega_X|^C |\Omega_Y|^C,$$

c'est-à-dire

$$(|\Omega_Y|^C + 2)(|\Omega_Y|^C + 1) \geq |\Omega_X|^C,$$

ce qui est évident puisqu'on suppose $|\Omega_X| \leq |\Omega_Y|$.

□

Le théorème démontre l'identifiabilité générique de A_{hmm} et ϕ_{hmm} mais ne démontre pas l'identifiabilité générique des paramètres de ϕ et A. Il nous faut démontrer que les paramètres associés aux probabilités du modèle sont bien génériquement identifiables.

4.6.1 Identifiabilité pour un modèle de loi Binomiale avec la fonction logistique

Il nous faut démontrer que pour tout (μ, ν), $(\mu', \nu') \in \mathbb{R}^4 \times \mathbb{R}^4$,

$$(A_{hmm}(\mu, \nu), \phi_{hmm}(\mu, \nu)) = (A_{hmm}(\mu', \nu'), \phi_{hmm}(\mu', \nu')) \Rightarrow (\mu, \nu) = (\mu', \nu').$$

Dans la représentation de la chaîne de Markov cachée, la matrice d'émission ϕ_{hmm} est déterministe et indépendante de μ, ν. Ainsi, il nous faut montrer que A_{hmm} est injective. On rappelle que $A_{hmm}(\mu,\nu)(h_n, h_{n-1}) = \mathbb{P}(h_n|h_{n-1},\nu,\mu)$. La probabilité de transition de la chaîne de Markov est :

$$\begin{aligned}
\mathbb{P}(h_n|h_{n-1},\mu,\nu) =& \mathbb{P}(x_n^C, y_{n+1}^C|x_{n-1}^C, y_n^C) \\
=& \mathbb{P}(y_{n+1}^C|x_n^C, y_n^C)\mathbb{P}(x_n^C|x_{n-1}^C, y_n^C) \\
=& \prod_{c=1}^{C} \mathbb{P}(y_{c,n+1}|x_n^C, y_n^C)\mathbb{P}(x_{c,n}|x_{c,n-1}, y_n^C) \\
=& \prod_{c=1}^{C} \binom{|\Omega_Y|-1}{y_{c,n+1}} \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\mu_1\frac{x_{c,n}}{|\Omega_X|}+\mu_2\frac{y_{c,n}}{|\Omega_Y|}+\mu_3\frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})}}\right]^{|\Omega_Y|-1} \\
& \times \left[\mathrm{e}^{-(\mu_0+\mu_1\frac{x_{c,n}}{|\Omega_X|}+\mu_2\frac{y_{c,n}}{|\Omega_Y|}+\mu_3\frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})}\right]^{|\Omega_Y|-y_{c,n+1}-1} \\
& \times \binom{|\Omega_X|-1}{x_{c,n}} \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_1\frac{x_{c,n-1}}{|\Omega_X|}+\nu_2\frac{y_{c,n}}{|\Omega_Y|}+\nu_3(\frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|}))}}\right]^{|\Omega_X|-1} \\
& \times \left[\mathrm{e}^{-(\nu_0+\nu_1\frac{x_{c,n-1}}{|\Omega_X|}+\nu_2\frac{y_{c,n}}{|\Omega_Y|}+\nu_3(\frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})}\right]^{|\Omega_X|-x_{c,n}-1}
\end{aligned}$$

où f est la fonction d'agrégation des champs voisins.

De cette expression, on établit le théorème suivant.

Théorème 12. *Les paramètres $(\mu,\nu) = (\mu_0,\mu_1,\mu_2,\mu_3,\nu_0,\nu_1,\nu_2,\nu_3)$ du MHMM-DF sont génériquement identifiables à partir de 7 observations consécutives si $|\Omega_X| \leq |\Omega_Y|$.*

Preuve.

Supposons que pour tout $(h_{n-1}, h_n) \in (|\Omega_X|^C \times |\Omega_Y)^C)^2$, on a

$$\mathbb{P}(h_n|h_{n-1},\mu,\nu) = \mathbb{P}(h_n|h_{n-1},\mu',\nu').$$

Montrons alors que $(\mu,\nu) = (\mu',\nu')$. On va montrer ceci coordonnée par coordonnée.

D'abord, montrons que $(\mu_0,\nu_0) = (\mu_0',\nu_0')$.

On pose

$$\begin{aligned}
L1(\mu_0,\nu_0) &= \mathbb{P}\left((\mathbf{0}^C,\mathbf{0}^C)|(\mathbf{0}^C,\mathbf{0}^C),\mu,\nu\right) \\
&= \binom{|\Omega_Y|-1}{0}^C \left[\frac{\mathrm{e}^{-\mu_0}}{1+\mathrm{e}^{-\mu_0}}\right]^{(|\Omega_Y|-1)C} \times \binom{|\Omega_X|-1}{0}^C \left[\frac{\mathrm{e}^{-\nu_0}}{1+\mathrm{e}^{-\nu_0}}\right]^{(|\Omega_X|-1)C} \\
&= \left[\frac{\mathrm{e}^{-\mu_0}}{1+\mathrm{e}^{-\mu_0}}\right]^{(|\Omega_Y|-1)C} \times \left[\frac{\mathrm{e}^{-\nu_0}}{1+\mathrm{e}^{-\nu_0}}\right]^{(|\Omega_X|-1)C}
\end{aligned}$$

et

$$
\begin{aligned}
L2(\mu_0,\nu_0) &= \mathbb{P}\left((\mathbf{0}^C,\mathbf{0}^C)|\left(\mathbf{0}^C,\begin{pmatrix}|\Omega_Y|-1\\0\\\vdots\\0\end{pmatrix}\right),\mu,\nu\right)\\
&= \binom{|\Omega_Y|-1}{0}^{C-1}\binom{|\Omega_Y|-1}{|\Omega_Y|-1}\left[\frac{\mathrm{e}^{-\mu_0(C-1)}}{\left(1+\mathrm{e}^{-\mu_0}\right)^C}\right]^{|\Omega_Y|-1}\times\binom{|\Omega_X|-1}{0}^{C}\left[\frac{\mathrm{e}^{-\nu_0}}{1+\mathrm{e}^{-\nu_0}}\right]^{(|\Omega_X|-1)C}\\
&= \left[\frac{\mathrm{e}^{-\mu_0(C-1)}}{\left(1+\mathrm{e}^{-\mu_0}\right)^C}\right]^{|\Omega_Y|-1}\times\left[\frac{\mathrm{e}^{-\nu_0}}{1+\mathrm{e}^{-\nu_0}}\right]^{(|\Omega_X|-1)C}
\end{aligned}
$$

Par hypothèse, on a $L1(\mu_0,\nu_0)=L1(\mu_0',\nu_0')$ et $L2(\mu_0,\nu_0)=L2(\mu_0',\nu_0')$. Ainsi alors $L1(\mu_0,\nu_0)/L2(\mu_0,\nu_0)=L1(\mu_0',\nu_0')/L2(\mu_0',\nu_0')$, ce qui implique que $e^{-(|\Omega_Y|-1)\mu_0}=e^{-(|\Omega_Y|-1)\mu_0'}$. Par injectivité de l'exponentielle, on a donc $\mu_0'=\mu_0$. Puisque $L1(\mu_0,\nu_0)=L1(\mu_0,\nu_0')$, on a

$$
\left(\left(\frac{1}{1+e^{-\nu_0}}\right)e^{-\nu_0}\right)^{(|\Omega_X|-1)C}=\left(\left(\frac{1}{1+e^{-\nu_0'}}\right)e^{-\nu_0'}\right)^{(|\Omega_X|-1)C},
$$

ce qui est équivalent à

$$
\left(1+e^{-\nu_0}\right)e^{\nu_0}=\left(1+e^{-\nu_0'}\right)e^{\nu_0'}
$$

car $\left(1+e^{-\nu_0}\right)e^{\nu_0}$ et $\left(1+e^{-\nu_0'}\right)e^{\nu_0'}$ sont positifs. On en déduit donc que $\nu_0=\nu_0$.

Montrons que $(\mu_1,\nu_1)=(\mu_1',\nu_1')$. On pose

$$
\begin{aligned}
L3(\mu_0,\mu_1,\nu_0,) =&\mathbb{P}\left((\mathbf{0}^C,\mathbf{0}^C)|\left(\begin{pmatrix}|\Omega_X|-1\\0\\\vdots\\0\end{pmatrix},\mathbf{0}^C\right),\mu,\nu\right)\\
&\propto\left[\frac{1}{1+\mathrm{e}^{-(\mu_0)}}\right]^{(|\Omega_Y|-1)(C-1)}\left[\mathrm{e}^{-(\mu_0)}\right]^{(|\Omega_Y|-1)(C-1)}\\
&\quad\times\left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1)}}\right]^{(|\Omega_Y|-1)}\left[\mathrm{e}^{-(\mu_0+\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1)}\right]^{(|\Omega_Y|-1)}\\
&\quad\times\left[\frac{1}{1+\mathrm{e}^{-(\nu_0)}}\right]^{(|\Omega_X|-1)C}\left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-1)(C-1)}
\end{aligned}
$$

Par hypothèse $L3(\mu_0,\mu_1,\nu_0,)=L3(\mu_0,\mu_1',\nu_0,)$, ce qui est équivalent à

$$
\left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1')}}\right]\mathrm{e}^{\frac{-(|\Omega_X|-1)}{|\Omega_X|}\mu_1'}=\left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1)}}\right]\mathrm{e}^{-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1}
$$

D'où

$$
\mathrm{e}^{-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1}+\mathrm{e}^{-\mu_0-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1'}=\mathrm{e}^{-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1'}+\mathrm{e}^{-\mu_0-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1-\frac{(|\Omega_X|-1)}{|\Omega_X|}\mu_1'},
$$

ce qui implique $\mu_1=\mu_1'$. Maintenant, posons

$$
\begin{aligned}
L4(\mu_0,\nu_0,\nu_1) =& \mathbb{P}\left(\left((|\Omega_X|-1)\times \mathbf{1}^C, \mathbf{0}^C\right)|(\mathbf{0}^C,\mathbf{0}^C),\mu,\nu\right)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\mu_0)}}\right]^{(|\Omega_Y|-1)(C-1)} \left[\mathrm{e}^{-(\mu_0)}\right]^{(|\Omega_Y|-1)(C-1)}\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\mu_0)}}\right]^{(|\Omega_Y|-1)C} \left[\mathrm{e}^{-(\mu_0)}\right]^{(|\Omega_Y|-1)C}\\
&\quad\times \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_1\frac{(|\Omega_X|-1)}{|\Omega_X|})}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0+\nu_1\frac{(|\Omega_X|-1)}{|\Omega_X|})}\right]^{(|\Omega_X|-1)C}
\end{aligned}
$$

Puisque $L4(\mu_0,\nu_0,\nu_1') = L4(\mu_0,\nu_0,\nu_1)$, on a

$$
\left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_1'\frac{(|\Omega_X|-1)}{|\Omega_X|})}}\right]\left[\mathrm{e}^{-\nu_1'\frac{(|\Omega_X|-1)}{|\Omega_X|}}\right] = \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_1\frac{(|\Omega_X|-1)}{|\Omega_X|})}}\right]\left[\mathrm{e}^{-\nu_1\frac{(|\Omega_X|-1)}{|\Omega_X|}}\right],
$$

ce qui donne $\nu_1' = \nu_1$.

Reste à montrer que les paramètres ν_2, ν_3, μ_2 et μ_3 sont identifiables. Cependant, la preuve dépend de la fonction d'agrégation choisie. Ainsi deux démonstrations sont faites, la première avec l'agrégation moyennée et la seconde avec l'agrégation alphabétique.

4.6.1.1 Identifiabilité pour l'agrégation moyennée

Montrons l'identifiabilité de μ_2 et ν_2 dans le cas où $f = f_1$. Puisque $f_1\left(\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}\right) = 0$, on pose

$$
\begin{aligned}
L5(\mu_2,\nu_0,\nu_2) =& \mathbb{P}\left(\left(\mathbf{0}^C, \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}\right)|(\mathbf{0}^C,\mathbf{0}^C),\mu,\nu\right)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2)}}\right]^{(|\Omega_X|-1)} \left[\frac{1}{1+\mathrm{e}^{-\nu_0}}\right]^{(|\Omega_X|-1)(C-1)}\\
&\quad\times \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2)}\right]^{(|\Omega_X|-1)} \left[\mathrm{e}^{-\nu_0}\right]^{(|\Omega_X|-1)(C-1)}\\
&\quad\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}}\right]^{(|\Omega_Y|-1)} \left[\frac{1}{1+\mathrm{e}^{-\mu_0}}\right]^{(|\Omega_Y|-1)(C-1)}\\
&\quad\times \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}\right]^{(|\Omega_Y|-1)} \left[\mathrm{e}^{-\mu_0}\right]^{(|\Omega_Y|-1)(C-1)}
\end{aligned}
$$

et

$$
\begin{aligned}
L6(\mu_0,\mu_2,\nu_0,\nu_2) =& \mathbb{P}\left(\left(\begin{pmatrix}\mathbf{0}^C,\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}\end{pmatrix} \mid \left(\mathbf{0}^C,\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}\right),\mu,\nu\right)\right.\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2)}}\right]^{(|\Omega_X|-1)} \left[\frac{1}{1+\mathrm{e}^{-(\nu_0)}}\right]^{(|\Omega_X|-1)(C-1)}\\
&\times \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2)}\right]^{(|\Omega_X|-1)} \left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-1)(C-1)}\\
&\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}}\right]^{(|\Omega_Y|-1)} \left[\frac{1}{1+\mathrm{e}^{-(\mu_0)}}\right]^{(|\Omega_Y|-1)(C-1)}\\
&\times \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}\right]^{(|\Omega_Y|-2)} \left[\mathrm{e}^{-(\mu_0)}\right]^{(|\Omega_Y|-1)(C-1)}
\end{aligned}
$$

Alors

$$\frac{L5(\mu_2,\nu_0,\nu_2)}{L6(\mu_2,\mu_2,\nu_0,\nu_2)} \propto \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}\right]$$

Puisque $L5(\mu_2,\nu_0,\nu_2) = L5(\mu_2',\nu_0,\nu_2')$ et $L6(\mu_0,\mu_2,\nu_0,\nu_2) = L6(\mu_0,\mu_2',\nu_0,\nu_2')$, on en déduit que

$$\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)} = \mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2')},$$

ce qui montre que $\mu_2 = \mu_2'$.

Puisque μ_2 est identifiable et $L5(\mu_0,\mu_2,\nu_0,\nu_2) = L5(\mu_0,\mu_2,\nu_0,\nu_2')$, on en déduit que

$$\mathrm{e}^{\frac{1}{|\Omega_Y|}\nu_2} = \mathrm{e}^{\frac{1}{|\Omega_Y|}\nu_2'}$$

et donc $\nu_2 = \nu_2'$.

Désormais montrons l'identifiabilité des paramètres μ_3 et ν_3. On pose

$$
\begin{aligned}
L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{1}^C\right) \mid \left(\mathbf{0}^C,\mathbf{0}^C\right),\mu,\nu\right)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}\right]^{(|\Omega_X|-1)C}\\
&\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}}\right]^{(|\Omega_Y|-1)C} \times \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}\right]^{(|\Omega_Y|-1)C}
\end{aligned}
$$

et

$$
\begin{aligned}
L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}((\mathbf{0}^C,\mathbf{1}^C)|(\mathbf{0}^C,\mathbf{1}^C),\mu,\nu)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}\right]^{(|\Omega_X|-1)C}\\
&\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}}\right]^{(|\Omega_Y|-1)C} \times \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}\right]^{(|\Omega_Y|-2)C}
\end{aligned}
$$

Par hypothèse, on a

$$\frac{L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}{L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)} = \frac{L7(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3')}{L8(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3')}$$

avec

$$\frac{L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}{L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)} \propto \quad \mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}.$$

Par injectivité de l'exponentielle, on en déduit $\mu_3 = \mu_3'$. Enfin, sachant que

$$L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) \propto \mathrm{e}^{\frac{1}{|\Omega_Y|}\nu_3}$$

et que $L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) = L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3')$, on montre facilement que $\nu_3 = \nu_3'$.

□

4.6.1.2 Identifiabilité pour l'agrégation alphabétique

Pour prouver l'identifiabilité de μ_2, μ_3, ν_2 et ν_3 quand $f = f_2$. Posons

$$\begin{aligned}
L5(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}((\mathbf{0}^C,\mathbf{1}^C)|(\mathbf{0}^C,\mathbf{0}^C),\mu,\nu) \\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\nu_3)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\nu_3)}\right]^{(|\Omega_X|-1)C} \\
&\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3)}}\right]^{(|\Omega_Y|-1)C} \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3)}\right]^{(|\Omega_Y|-1)C}
\end{aligned}$$

et

$$\begin{aligned}
L6(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}((\mathbf{0}^C,\mathbf{1}^C)|(\mathbf{0}^C,(|\Omega_Y|-1)\times\mathbf{1}^C),\mu,\nu) \\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\nu_3)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\nu_3)}\right]^{(|\Omega_X|-1)C} \\
&\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3)}}\right]^{(|\Omega_Y|-1)C}
\end{aligned}$$

Alors

$$\frac{L5(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}{L6(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)} \propto \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3)}\right]^{(|\Omega_Y|-1)C}$$

puisque $L5(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) = L5(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3')$ et $L6(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) = L6(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3')$, on en déduit que

$$\left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3)}\right]^{(|\Omega_Y|-1)C} = \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2'+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3')}\right]^{(|\Omega_Y|-1)C}$$

Par conséquent,

$$\frac{1}{|\Omega_Y|}\mu_2 + \frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3 = \frac{1}{|\Omega_Y|}\mu_2' + \frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3'.$$

Maintenant il nous reste à utiliser deux autres équations afin de pouvoir prouver l'identifiabilité des paramètres. On pose

$$L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)=\mathbb{P}((\mathbf{0}^C,\mathbf{2}^C)|(\mathbf{0}^C,\mathbf{0}^C),\mu,\nu)$$

et

$$L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)=\mathbb{P}((\mathbf{0}^C,\mathbf{2}^C)|(\mathbf{0}^C,(|\Omega_Y|-1)\times\mathbf{1}^C),\mu,\nu).$$

Alors

$$\frac{L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}{L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}\propto\frac{2}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3.$$

Puisque $L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)=L7(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3')$ et $L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)=L8(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3')$ on en déduit que

$$\frac{2}{|\Omega_Y|}\mu_2+\frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\mu_3=\frac{2}{|\Omega_Y|}\mu_2'+\frac{f_2(2,..,2)}{|f_2(\Omega_Y^{C-1})|}\mu_3'$$

En soustrayant $\frac{L7(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}{L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}-2\frac{L5(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}{L6(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3)}$ on obtient

$$\frac{f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3=\frac{f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3'.$$

Donc si $f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)\neq 0$, l'identifiabilité de μ_3 est vérifiée. On rappelle la définition de la fonction d'agrégation alphabétique f_2 :

$$f_2(Y_n^{\mathcal{C}\setminus c})=\sum_{c=C-1}^{1}\mathbb{1}_{g_c(Y_n^{\mathcal{C}\setminus c})>g_{c+1}(Y_n^{\mathcal{C}\setminus c})\}}\sum_{j=g_{c+1}(Y_n^{\mathcal{C}\setminus c})+1}^{g_c(Y_n^{\mathcal{C}\setminus c})}\binom{|\Omega_Y|-j+c-1}{c-1}$$

Par conséquent

$$f_2(\mathbf{1}^C)=\binom{|\Omega_Y|+C-3}{C-2}$$

$$f_2(\mathbf{2}^C)=\sum_{j=1}^{2}\binom{|\Omega_Y|-j+C-2}{C-2}$$

Ainsi

$$\begin{array}{rcl}f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C) & = & -2\binom{|\Omega_Y|+C-3}{C-2}+\sum_{j=1}^{2}\binom{|\Omega_Y|-j+C-2}{C-2}\\ & = & \binom{|\Omega_Y|+C-4}{C-2}-\binom{|\Omega_Y|+C-3}{C-2}\\ & = & \frac{(|\Omega_Y|+C-4)!}{(|\Omega_Y|-2)!(C-2)!}-\frac{(|\Omega_Y|+C-3)!}{(|\Omega_Y|-1)!(C-2)!}\\ & = & \frac{(|\Omega_Y|+C-4)!}{(|\Omega_Y|-2)!(C-2)!}(1-\frac{|\Omega_Y|+C-3}{|\Omega_Y|-1})\end{array}$$

Montrons que $f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)=0$ implique $C=2$. puisque $\frac{(|\Omega_Y|+C-4)!}{(|\Omega_Y|-2)!(C-2)!}>0$, si $f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)=0$ alors $1-\frac{|\Omega_Y|+C-3}{|\Omega_Y|-1}=0$. Donc

$$\begin{array}{rcl}1 & = & \frac{|\Omega_Y|+C-3}{|\Omega_Y|-1}\\ |\Omega_Y|-1 & = & |\Omega_Y|+C-3\\ C & = & 2\end{array}$$

Puisque $f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)=0$ que quand $C=2$ alors quand $C\geq 3$ on a $\mu_3=\mu_3'$. Finalement, en utilisant l'équation $\frac{L5}{L6}$ on obtient facilement que si $\mu_3=\mu_3'$ alors $\mu_2=\mu_2'$.

Il nous reste à montrer que ν_2 et ν_3 sont identifiables. Puisque μ_2 et μ_3 sont identifiables en reprenant les équations $L5$ et $L6$ on peut montrer l'identifiabilité de ν_2 et ν_3. Puisque $L5(\mu_0, \mu_2, \mu_3, \nu_0, \nu_2, \nu_3) = L5(\mu_0, \mu_2, \mu_3, \nu_0, \nu_2', \nu_3')$, on en déduit que

$$\nu_0 + \frac{1}{|\Omega_Y|}\nu_2 + \frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\nu_3 = \nu_0 + \frac{1}{|\Omega_Y|}\nu_2' + \frac{f_2(1,..,1)}{|f_2(\Omega_Y^{C-1})|}\nu_3'$$

De façon analogue puisque $L6(\mu_0, \mu_2, \mu_3, \nu_0, \nu_2, \nu_3) = L6(\mu_0, \mu_2, \mu_3, \nu_0, \nu_2', \nu_3')$ on en déduit que

$$\nu_0 + \frac{2}{|\Omega_Y|}\nu_2 + \frac{f_2(2,..,2)}{|f_2(\Omega_Y^{C-1})|}\nu_3 = \nu_0 + \frac{2}{|\Omega_Y|}\nu_2' + \frac{f_2(2,..,2)}{|f_2(\Omega_Y^{C-1})|}\nu_3'$$

Ainsi en appliquant la même méthode que précédemment pour μ_2 et μ_3 on peut prendre $L6 - 2L5$ pour obtenir que ν_3 et ν_2 sont identifiables si $C > 2$.

4.6.2 Identifiabilité du MHMM-DF avec loi d'émission BU et loi de transition BL

Il nous faut démontrer que pour tout $(\mu, \nu), (\mu', \nu') \in \mathbb{R}^4 \times \mathbb{R}^4$

$$(A_{hmm}(\mu, \nu), \phi_{hmm}(\mu, \nu)) = (A_{hmm}(\mu', \nu'), \phi_{hmm}(\mu', \nu') \Rightarrow (\mu, \nu) = (\mu', \nu')$$

Dans la représentation de la chaîne de Markov cachée, la matrice d'émission ϕ_{hmm} est déterministe et indépendante de μ et ν. Ainsi, il nous faut montrer que $A_{hmm}(\mu, \nu) = A_{hmm}(\mu', \nu')$ implique $(\mu, \nu) = (\mu', \nu')$. On rappelle que $A_{hmm}(\mu, \nu)(h_n, h_{n_1}) = \mathbb{P}(h_n | h_{n-1}, \nu, \mu)$. La probabilité de transition de la chaîne de Markov est

$$\begin{aligned}
\mathbb{P}(h_n | h_{n-1}, \mu, \nu) &= \mathbb{P}(x_n^C, y_{n+1}^C | x_{n-1}^C, y_n^C) \\
&= \mathbb{P}(y_{n+1}^C | x_n^C, y_n^C) \mathbb{P}(x_n^C | x_{n-1}^C, y_n^C) \\
&= \prod_{c=1}^{C} \mathbb{P}(y_{c,n+1} | x_n^C, y_n^C) \mathbb{P}(x_{c,n} | x_{c,n-1}, y_n^C) \\
&= \prod_{c=1}^{C} \left(\mathbb{1}_{\{x_{c,n}=0 \cap y_{c,n+1}=0\}} + \mathbb{1}_{\{x_{c,n}\neq 0\}} \binom{|\Omega_Y|-1}{y_{c,n+1}} \left[\frac{1}{1+\mathrm{e}^{-(\mu_0 + \mu_1 \frac{x_{c,n}}{|\Omega_X|} + \mu_2 \frac{y_{c,n}}{|\Omega_Y|} + \mu_3 \frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})}} \right]^{|\Omega_Y|-1} \right. \\
&\quad \left. \times \left[\mathrm{e}^{-(\mu_0 + \mu_1 \frac{x_{c,n}}{|\Omega_X|} + \mu_2 \frac{y_{c,n}}{|\Omega_Y|} + \mu_3 \frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})} \right]^{|\Omega_Y|-y_{c,n+1}-1} \right) \\
&\quad \times \binom{|\Omega_X|-1}{x_{c,n}} \left[\frac{1}{1+\mathrm{e}^{-(\nu_0 + \nu_1 \frac{x_{c,n-1}}{|\Omega_X|} + \nu_2 \frac{y_{c,n}}{|\Omega_Y|} + \nu_3 (\frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|}))}} \right]^{|\Omega_X|-1} \\
&\quad \times \left[\mathrm{e}^{-(\nu_0 + \nu_1 \frac{x_{c,n-1}}{|\Omega_X|} + \nu_2 \frac{y_{c,n}}{|\Omega_Y|} + \nu_3 (\frac{f(y_n^{\mathcal{C}\setminus c})}{|f(\Omega_Y^{C-1})|})} \right]^{|\Omega_X|-x_{c,n}-1}
\end{aligned}$$

où f est la fonction d'agrégation des champs voisins. De cette expression, on établit le théorème suivant.

Théorème 13. *Les paramètres $(\mu, \nu) = (\mu_0, \mu_1, \mu_2, \mu_3, \nu_0, \nu_1, \nu_2, \nu_3)$ du MHMM-DF sont génériquement identifiables à partir de 7 observations consécutives si $|\Omega_X| \leq |\Omega_Y|$ et $C > 2$.*

Preuve. Supposons que pour tout $(h_{n-1}, h_n) \in (|\Omega_X|^C, |\Omega_Y^C)^2$ on a

$$\mathbb{P}(h_n|h_{n-1}, \mu, \nu) = \mathbb{P}(h_n|h_{n-1}, \mu', \nu')$$

Montrons alors que $(\mu, \nu) = (\mu', \nu')$. Nous allons montrer ceci coordonnée par coordonnée. Commençons par montrer que μ_0 est génériquement identifiable. On pose

$$\begin{aligned} L1(\nu_0) =& \mathbb{P}((\mathbf{0}^C, \mathbf{0}^C)|(\mathbf{0}^C, \mathbf{0}^C), \mu, \nu) \\ =& \binom{|\Omega_X| - 1}{0}^C \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-1)C} \end{aligned}$$

Puisque $L1(\nu_0) = L1(\nu_0')$, on a

$$\left[\frac{1}{1 + \mathrm{e}^{-(\nu_0)}}\right] \left[\mathrm{e}^{-(\nu_0)}\right] = \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0')}}\right] \left[\mathrm{e}^{-(\nu_0')}\right]$$

De ce fait, $\nu_0 = \nu_0'$.

Nous avons montré l'identifiabilité de ν_0. Montrons l'identifiabilité de μ_0 et μ_1 à l'aide de deux équations. On pose

$$\begin{aligned} L2(\mu_0, \mu_1) =& \mathbb{P}\left((\mathbf{0}^C, \mathbf{0}^C)| \left(\mathbf{0}^C, \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right), \mu, \nu\right) \\ =& \binom{|\Omega_Y| - 1}{0} \left[\frac{1}{1 + \mathrm{e}^{-(\mu_0 + \frac{\mu_1}{|\Omega_X|})}}\right]^{(|\Omega_Y|-1)} \left[\mathrm{e}^{-(\mu_0 + \frac{\mu_1}{|\Omega_X|})}\right]^{(|\Omega_Y|-1)} \\ & \times \binom{|\Omega_X| - 1}{0}^{(C-1)} \binom{|\Omega_X| - 1}{1} \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-1)(C-1)} \left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-2)} \end{aligned}$$

Puisque $L2(\mu_0, \mu_1) = L2(\mu_0', \mu_1')$ on en déduit que

$$\left[\frac{1}{1 + \mathrm{e}^{-(\mu_0 + \frac{\mu_1}{|\Omega_X|})}}\right]^{(|\Omega_Y|-1)} \left[\mathrm{e}^{-(\mu_0 + \frac{\mu_1}{|\Omega_X|})}\right]^{(|\Omega_Y|-1)} = \left[\frac{1}{1 + \mathrm{e}^{-(\mu_0' + \frac{\mu_1'}{|\Omega_X|})}}\right]^{(|\Omega_Y|-1)} \left[\mathrm{e}^{-(\mu_0' + \frac{\mu_1'}{|\Omega_X|})}\right]^{(|\Omega_Y|-1)},$$

ce qui est équivalent à

$$\frac{\mathrm{e}^{-(\mu_0 + \frac{\mu_1}{|\Omega_X|})}}{1 + \mathrm{e}^{-(\mu_0 + \frac{\mu_1}{|\Omega_X|})}} = \frac{\mathrm{e}^{-(\mu_0' + \frac{\mu_1'}{|\Omega_X|})}}{1 + \mathrm{e}^{-(\mu_0' + \frac{\mu_1'}{|\Omega_X|})}}$$

qui est aussi équivalent à

$$\mu_0 + \frac{\mu_1}{|\Omega_X|} = \mu_0' + \frac{\mu_1'}{|\Omega_X|}.$$

Afin de montrer l'identifiabilité des paramètres μ_0 et μ_1 on utilise aussi une autre équation. On pose

$$L3(\mu_0, \mu_1) = \mathbb{P}\left((\mathbf{0}^C, \mathbf{0}^C)| \left(\mathbf{0}^C, \begin{pmatrix} |\Omega_X| - 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right), \mu, \nu\right)$$

Puisque $L3(\mu_0, \mu_1) = L3(\mu'_0, \mu'_1)$ on en déduit que

$$\mu_0 + \frac{\mu_1(|\Omega_X| - 1)}{|\Omega_X|} = \mu'_0 + \frac{\mu'_1(|\Omega_X| - 1)}{|\Omega_X|}$$

Par hypothèse, on a $L2(\mu_0, \mu_1) = L2(\mu'_0, \mu'_1)$ et $L3(\mu_0, \mu_1) = L3(\mu'_0, \mu'_1)$. Ainsi $L3(\mu_0, \mu_1) - L2(\mu_0, \mu_1) = L3(\mu'_0, \mu'_1) - L2(\mu'_0, \mu'_1)$. On en déduit que

$$\frac{\mu_1(|\Omega_X| - 2)}{|\Omega_X|} = \frac{\mu'_1(|\Omega_X| - 2)}{|\Omega_X|}$$

Il en découle que $\mu_1 = \mu'_1$. Puisque $\mu_1 = \mu'_1$, à l'aide de l'équation $L2$ on obtient $\mu_0 = \mu'_0$.
Montrons l'identifiabilité de ν_1. On pose

$$L4(\mu_0, \nu_0, \nu_1) = \mathbb{P}\left(((|\Omega_X| - 1) \times \mathbf{1}^C, \mathbf{0}^C) | (\mathbf{0}^C, \mathbf{0}^C), \mu, \nu\right)$$

Ainsi,

$$L4(\mu_0, \nu_0, \nu_1) \propto \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0 + \nu_1 \frac{(|\Omega_X|-1)}{|\Omega_X|})}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0 + \nu_1 \frac{(|\Omega_X|-1)}{|\Omega_X|})}\right]^{(|\Omega_X|-1)C}$$

Puisque $L4(\mu_0, \nu_0, \nu_1) = L4(\mu_0, \nu_0, \nu'_1)$, on a

$$\left[\frac{1}{1 + \mathrm{e}^{-(\nu_0 + \nu_1 \frac{(|\Omega_X|-1)}{|\Omega_X|})}}\right] \left[\mathrm{e}^{-\nu_1 \frac{(|\Omega_X|-1)}{|\Omega_X|}}\right] = \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0 + \nu'_1 \frac{(|\Omega_X|-1)}{|\Omega_X|})}}\right] \left[\mathrm{e}^{-\nu'_1 \frac{(|\Omega_X|-1)}{|\Omega_X|}}\right]$$

On en déduit que $\nu_1 = \nu'_1$.

Nous avons montré que tous les paramètres sont identifiables sauf pour ν_2, ν_3, μ_2 et μ_3. Comme précédemment, on donnera deux démonstrations, selon la fonction d'agrégation.

4.6.2.1 Identifiabilité pour l'agrégation moyennée

Montrons l'identifiabilité et ν_2. On pose

$$\begin{aligned} L5(\nu_0, \nu_2) =& \mathbb{P}\left(\left(\mathbf{0}^C, \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}\right) | (\mathbf{0}^C, \mathbf{0}^C), \mu, \nu\right) \\ \propto& \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0 + \frac{1}{|\Omega_Y|}\nu_2)}}\right]^{(|\Omega_X|-1)} \left[\frac{1}{1 + \mathrm{e}^{-(\nu_0)}}\right]^{(|\Omega_X|-1)(C-1)} \\ & \times \left[\mathrm{e}^{-(\nu_0 + \frac{1}{|\Omega_Y|}\nu_2)}\right]^{(|\Omega_X|-1)} \left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-1)(C-1)} \end{aligned}$$

Puisque $L5(\nu_0, \nu_2) = L5(\nu_0, \nu'_2)$, on a

$$\nu_0 + \frac{1}{|\Omega_Y|}\nu_2 = \nu_0 + \frac{1}{|\Omega_Y|}\nu'_2$$

Par conséquent, $\nu_2 = \nu'_2$.

Montrons maintenant l'identifiabilité du paramètre μ_2. On pose

$$\begin{aligned}
L6(\mu_0,\mu_2,\nu_0,\nu_2) =& \mathbb{P}\left(\left(\mathbf{0}^C,\begin{pmatrix}1\\0\\...\\0\end{pmatrix}\right) \mid \left(\begin{pmatrix}1\\0\\...\\0\end{pmatrix},\mathbf{0}^C\right),\mu,\nu\right)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2)}}\right]^{(|\Omega_X|-1)} \left[\frac{1}{1+\mathrm{e}^{-(\nu_0)}}\right]^{(|\Omega_X|-1)(C-1)}\\
&\quad\times \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2)}\right]^{(|\Omega_X|-2)} \left[\mathrm{e}^{-(\nu_0)}\right]^{(|\Omega_X|-1)(C-1)}\\
&\quad\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}}\right]^{(|\Omega_Y|-1)} \times \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}\right]^{(|\Omega_Y|-1)}
\end{aligned}$$

Par conséquent

Puisque $L6(\mu_0,\mu_2,\nu_0,\nu_2) = L6(\mu_0,\mu_2',\nu_0,\nu_2)$, on en déduit que

$$\left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}}\right]^{(|\Omega_Y|-1)} \times\left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2)}\right]^{(|\Omega_Y|-1)} = \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2')}}\right]^{(|\Omega_Y|-1)} \times\left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2')}\right]^{(|\Omega_Y|-1)}$$

Il en résulte que $\mu_2 = \mu_2'$.

Nous allons désormais montrer que le paramètre ν_3 est identifiable. On pose

$$\begin{aligned}
L7(\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{1}^C\right) \mid \left(\mathbf{0}^C,\mathbf{0}^C\right),\mu,\nu\right)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}}\right]^{(|\Omega_X|-1)C} \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}\right]^{(|\Omega_X|-1)C}
\end{aligned}$$

Puisque $L7(\nu_0,\nu_2,\nu_3) = L7(\nu_0,\nu_2,\nu_3')$ on obtient

$$\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3 = \nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3'$$

En simplifiant encore, on obtient que $\nu_3 = \nu_3'$. Il nous reste a montrer l'identifiabilité de μ_3. On pose

$$\begin{aligned}
L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{1}^C\right) \mid \left(\begin{pmatrix}1\\0\\...\\0\end{pmatrix},\mathbf{0}^C\right),\mu,\nu\right)\\
&\propto \left[\frac{1}{1+\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}}\right]^{(|\Omega_X|-1)C}\\
&\quad\times \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}\right]^{(|\Omega_X|-1)(C-1)} \left[\mathrm{e}^{-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{1}{|\Omega_Y|}\nu_3)}\right]^{(|\Omega_X|-2)}\\
&\quad\times \left[\frac{1}{1+\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}}\right]^{(|\Omega_Y|-1)} \times \left[\mathrm{e}^{-(\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3)}\right]^{(|\Omega_Y|-1)}
\end{aligned}$$

Puisque $L8(\mu_0,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) = L8(\mu_0,\mu_2,\mu_3',\nu_0,\nu_2,\nu_3)$ on en déduit que

$$\mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3 = \mu_0+\frac{1}{|\Omega_Y|}\mu_2+\frac{1}{|\Omega_Y|}\mu_3'$$

et donc $\mu_3 = \mu_3'$. Nous avons montré l'identifiabilité générique de tous les paramètres avec l'agrégation moyennée pour le deuxième modèle. □

4.6.2.2 Identifiabilité pour l'agrégation alphabétique

Pour prouver l'identifiabilité de ν_2 et ν_3, on définit les équations $L5$ et $L6$.

$$\begin{aligned} L5(\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{1}^C\right) \mid \left(\mathbf{0}^C,\mathbf{0}^C\right),\mu,\nu\right) \\ \propto & \left[\frac{1}{1+\mathrm{e}^{(-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))}}\right]^{(|\Omega_X|-1)C} \\ & \times \left[\mathrm{e}^{(-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))}\right]^{(|\Omega_X|-1)C} \end{aligned}$$

et

$$\begin{aligned} L6(\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{2}^C\right) \mid \left(\mathbf{0}^C,\mathbf{0}^C\right),\mu,\nu\right) \\ \propto & \left[\frac{1}{1+\mathrm{e}^{(-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))}}\right]^{(|\Omega_X|-1)C} \\ & \times \left[\mathrm{e}^{(-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))}\right]^{(|\Omega_X|-1)C} \end{aligned}$$

Sachant que $L5(\nu_0,\nu_2,\nu_3) = L5(\nu_0,\nu_2',\nu_3')$, on a

$$\frac{1}{|\Omega_Y|}\nu_2 + \frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3 = \frac{1}{|\Omega_Y|}\nu_2' + \frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3'$$

De plus, $L6(\nu_0,\nu_2,\nu_3) = L6(\nu_0,\nu_2',\nu_3')$, ce qui implique

$$\frac{2}{|\Omega_Y|}\nu_2 + \frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3 = \frac{2}{|\Omega_Y|}\nu_2' + \frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3'$$

Puisque $L5(\nu_0,\nu_2,\nu_3) = L5(\nu_0,\nu_2',\nu_3')$ et $L6(\nu_0,\nu_2,\nu_3) = L6(\nu_0,\nu_2',\nu_3')$ alors $L6(\nu_0,\nu_2,\nu_3) - 2L5(\nu_0,\nu_2,\nu_3) = L6(\nu_0,\nu_2',\nu_3') - 2L5(\nu_0,\nu_2',\nu_3')$. Par conséquent,

$$\frac{f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3 = \frac{f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3'$$

Donc si $f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C) \neq 0$, l'identifiabilité de ν_3 est vérifiée. On rappelle la définition de la fonction d'agrégation alphabétique f_2 :

$$f_2(Y_n^{\mathcal{C}\setminus c}) = \sum_{c=C-1}^{1} \mathbb{1}_{g_c(Y_n^{\mathcal{C}\setminus c})>g_{c+1}(Y_n^{\mathcal{C}\setminus c})\}} \sum_{j=g_{c+1}(Y_n^{\mathcal{C}\setminus c})+1}^{g_c(Y_n^{\mathcal{C}\setminus c})} \binom{|\Omega_Y| - j + c - 1}{c-1}$$

Donc

$$f_2(\mathbf{1}^C) = \binom{|\Omega_Y| + C - 3}{C-2}$$

$$f_2(\mathbf{2}^C) = \sum_{j=1}^{2} \binom{|\Omega_Y| - j + C - 2}{C-2}$$

et

$$\begin{array}{rcl} f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C) & = & -2\binom{|\Omega_Y|+C-3}{C-2} + \sum_{j=1}^{2}\binom{|\Omega_Y|-j+C-2}{C-2} \\ & = & \binom{|\Omega_Y|+C-4}{C-2} - \binom{|\Omega_Y|+C-3}{C-2} \\ & = & \frac{(|\Omega_Y|+C-4)!}{(|\Omega_Y|-2)!(C-2)!} - \frac{(|\Omega_Y|+C-3)!}{(|\Omega_Y|-1)!(C-2)!} \\ & = & \frac{(|\Omega_Y|+C-4)!}{(|\Omega_Y|-2)!(C-2)!}(1 - \frac{|\Omega_Y|+C-3}{|\Omega_Y|-1}) \end{array}$$

Montrons que $f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C) = 0$ implique $C = 2$. Puisque $\frac{(|\Omega_Y|+C-4)!}{(|\Omega_Y|-2)!(C-2)!} > 0$, si

$f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C) = 0$ alors $1 - \frac{|\Omega_Y|+C-3}{|\Omega_Y|-1} = 0$. Donc

$$\begin{array}{rcl} 1 & = & \frac{|\Omega_Y|+C-3}{|\Omega_Y|-1} \\ |\Omega_Y| - 1 & = & |\Omega_Y| + C - 3 \\ C & = & 2 \end{array}$$

Puisque $f_2(\mathbf{2}^C) - 2f_2(\mathbf{1}^C) = 0$ quand $C = 2$, on a $\nu_3 = \nu_3'$ quand $C \geq 3$. Finalement, en utilisant l'équation $L6$ ou $L5$, on obtient facilement que si $\nu_3 = \nu_3'$ alors $\nu_2 = \nu_2'$. Il nous reste à montrer que μ_2 et μ_3 sont identifiables. On pose

$$\begin{aligned} L7(\mu_0,\mu_1,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{1}^C\right) \mid \left(\mathbf{1}^C,\mathbf{0}^C\right),\mu,\nu\right) \\ \propto& \left[\frac{1}{1+\mathrm{e}(-(\mu_0 + \frac{1}{|\Omega_X|}\mu_1 + \frac{1}{|\Omega_Y|}\mu_2 + \frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3))}\right]^{(|\Omega_X|-1)C} \\ &\times \left[\mathrm{e}(-(\mu_0 + \frac{1}{|\Omega_X|}\mu_1 + \frac{1}{|\Omega_Y|}\mu_2 + \frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3))\right]^{(|\Omega_X|-1)C} \\ &\times \left[\frac{1}{1+\mathrm{e}(-(\nu_0 + \frac{1}{|\Omega_Y|}\nu_2 + \frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))}\right]^{(|\Omega_X|-1)C} \\ &\times \left[\mathrm{e}(-(\nu_0 + \frac{1}{|\Omega_Y|}\nu_2 + \frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))\right]^{(|\Omega_X|-2)C} \end{aligned}$$

et

$$
\begin{aligned}
L8(\mu_0,\mu_1,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) =& \mathbb{P}\left(\left(\mathbf{0}^C,\mathbf{2}^C\right) \mid \left(\mathbf{1}^C,\mathbf{0}^C\right),\mu,\nu\right) \\
&\propto \left[\frac{1}{1+\mathrm{e}(-(\mu_0+\frac{1}{|\Omega_X|}\mu_1+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3))}\right]^{(|\Omega_X|-1)C} \\
&\times \left[\mathrm{e}(-(\mu_0+\frac{1}{|\Omega_X|}\mu_1+\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3))\right]^{(|\Omega_X|-1)C} \\
&\times \left[\frac{1}{1+\mathrm{e}(-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))}\right]^{(|\Omega_X|-1)C} \\
&\times \left[\mathrm{e}(-(\nu_0+\frac{1}{|\Omega_Y|}\nu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\nu_3))\right]^{(|\Omega_X|-2)C}
\end{aligned}
$$

Puisque $L7(\mu_0,\mu_1,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) = L7(\mu_0,\mu_1,\mu_2',\mu_3',\nu_0,\nu_2,\nu_3)$ et $L8(\mu_0,\mu_1,\mu_2,\mu_3,\nu_0,\nu_2,\nu_3) = L8(\mu_0,\mu_1,\mu_2',\mu_3',\nu_0,\nu_2,\nu_3)$, les trois prochaines équations sont vérifiées

$$
\begin{aligned}
\frac{1}{|\Omega_Y|}\mu_2+\frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3 &= \frac{1}{|\Omega_Y|}\mu_2'+\frac{f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3' \\
\frac{2}{|\Omega_Y|}\mu_2+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3 &= \frac{2}{|\Omega_Y|}\mu_2'+\frac{f_2(\mathbf{2}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3' \\
[L8-2L7](\nu_0,\nu_2,\nu_3,\mu_0,\mu_1,\mu_2,\mu_3) &= [L8-2L7](\nu_0,\nu_2,\nu_3,\mu_0,\mu_1,\mu_2',\mu_3')
\end{aligned}
$$

Ainsi on peut en déduire d'après $L8-2L7$ que

$$
\frac{f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3 = \frac{f_2(\mathbf{2}^C)-2f_2(\mathbf{1}^C)}{|f_2(\Omega_Y^{C-1})|}\mu_3'
$$

Comme précédemment pour les équations $L5$ et $L6$, on obtient que l'identifiabilité des paramètres μ_2 et μ_3 est vérifiée si $C>2$.

Nous avons montré l'identifiabilité générique de tous les paramètres avec l'agrégation alphabétique pour le deuxième modèle. □

4.7 Discussion des différentes paramétrisations

4.7.1 Les différentes paramétrisations de la loi d'émission

Le nombre de paramètres pour chaque modélisation de la loi d'émission ϕ est donné dans le tableau ci-dessous.

	Nombre de paramètres				
Modélisation	non paramétrique	Bin	BL	ZIBL	BU
ϕ	$\lvert\Omega_X\rvert^2\lvert\Omega_Y\rvert^C$	$\lvert\Omega_X\rvert\lvert\Omega_Y\rvert^C$	4	5	4
$\phi_{\text{Adventices}}$	$\lvert\Omega_X\rvert^2$	$\lvert\Omega_X\rvert$	2	3	2

Les deux modèles qui nécessitent le moins de paramètres sont les modèles binomiaux avec la fonction logistique avec ou sans l'hypothèse de la non-extinction de la variable cachée.

Paramétrer la probabilité d'émission à l'aide d'une loi Binomiale implique que tous les états de $Y_{c,n+1}$ peuvent être visités quand la probabilité de succès est différente de 0 ou 1. La probabilité de succès de la loi Binomiale logistique n'atteint jamais 0 ou 1 à part dans des cas dégénérés, ce qui pose problème pour modéliser la dynamique des adventices. En effet, dans le cas des adventices, si la banque de graines est vide alors il ne devrait pas être possible d'obtenir de la flore levée. Le processus de germination est difficile à modéliser à cause de la forte absence de flore levée dans les patchs. La loi Zero inflated Binomiale permet de résoudre ce problème en augmentant la probabilité d'obtenir une flore levée absente dans un patch. Cependant, la modélisation ZIB ne permet pas d'imposer une flore levée absente quand la banque de graines est absente. La modélisation Binomiale logistique avec non-extinction de la population cachée est une généralisation de la modélisation zero inflated Binomiale qui résout ce problème tout en conservant l'augmentation de la probabilité d'obtenir une flore levée absente dans un patch. L'utilisation d'autres lois comme la logistique cumulée ou bien la Beta cumulée aurait pu être envisagée (Herpigny and Gosselin, 2015). Cependant, la loi logistique cumulée a un nombre de paramètres dépendant du nombre d'états et ne permet pas de controler l'erreur-type (Herpigny and Gosselin, 2015). En ce qui concerne la loi Beta cumulée, elle nécessite d'avoir défini les bornes des classes. Même si la loi logistique cumulée et la loi Beta cumulée présentent de légers incovénients, il serait intéréssant de les utiliser au sein du modèle. En revanche, il faudrait montrer qu'un modèle utilisant l'une de ces deux lois est identifiable et que l'estimation via un tel modèle reste possible.

4.7.2 Les différentes paramétrisations de la loi de transition

Le nombre de paramètres pour chaque modélisation de la loi de transition A est donné dans le tableau ci-dessous.

	Nombre de paramètres			
Modélisation	Sparse	BL	PBL	BSSA
A	$(\lvert\Omega_X\rvert + 1) + 3$	4	4	5

Plusieurs points sont à discuter concernant la modélisation de A. Tout d'abord, la modélisation sparse est utile car elle permet d'obtenir des matrices de transition remplies de zéros. En contrepartie, cela impose une hypothèse de restriction sur la dynamique de la variable cachée. En effet, la variable cachée au temps $n+1$ ne peut prendre que trois valeurs x_n, x_n+1, x_n-1. De plus, le modèle sparse, utilisant des fonctions minimum et maximum, n'est pas dérivable en certains points, ce qui est très problématique lors de la mise à jour des paramètres via descente de gradient dans l'étape M.

La modélisation BSSA ainsi que la modélisation PBL limite l'influence de la variable cachée sur la variable cachée au temps suivant. Pour les adventices, l'état de la banque de graines dépend du nombre de graines qui ont survécu depuis l'année précédente dans la banque de graines et du nombre de graines rentrantes. Ces deux processus s'additionnent et génèrent l'état de la banque de graines au temps suivant. Afin de modéliser la dynamique des adventices, la modélisation BSSA ne paraît pas adéquate car elle suppose que l'état de la banque de graines au temps suivant dépend soit du processus de survie soit du processus de grenaison locale. La modélisation BL, qui elle au contraire ne modélise pas séparément ces deux processus, facilite leur comparaison. La modélisation PBL, quant à elle, limite l'influence de la survie des graines par rapport aux autres processus sur l'état de la banque de graines au temps suivant. La plupart des paramétrisations étudiées dans ce chapitre reposent sur l'utilisation de la fonction logistique car l'ensemble image de la fonction logistique est $[0, 1]$ et permet de facilement modéliser des probabilités. Cependant il se pourrait que la fonction logistique, et surtout l'influence des processus sous la forme d'une combinaison linéaire, ne soit pas pertinente pour modéliser la dynamique des adventices.

Chapitre 5

Estimation des paramètres d'un MHMM-DF

L'inférence du MHMM-DF est réalisée par l'algorithme EM (Dempster et al., 1977 ; Rabiner and Juang, 1986). Comme décrit dans la section 2.3.3.2, l'algorithme EM itère deux étapes :l'étape E qui utilise l'algorithme du Forward-backward (Rabiner, 1989 ; Rabiner and Juang, 1986) et l'étape M qui actualise les paramètres du modèle. Deux problèmes sont apparus lors de la mise en oeuvre de l'algorithme EM pour le MHMM-DF complet, tous deux en rapport avec les équations du Forward-Backward. Tout d'abord, le Forward-Backward ne peut être appliqué à toutes les chaînes à la fois sur le MHMM-DF complet car ses équations reposent sur l'indépendance des variables observées sachant les variables cachées. En revanche, en supprimant les dépendances entre variables observées, c'est-à-dire en se focalisant uniquement sur le sous-graphe du MHMM-DF associé à la dynamique des adventices, le Forward-Backward peut être employé sur toutes les chaînes à la fois. Cependant, cela rend sa complexité algorithmique exponentielle en le nombre de chaînes. La complexité algorithmique peut être améliorée jusqu'à être linéaire en le nombre de chaînes en appliquant le Forward-Backward chaîne par chaîne, et cela même pour le graphe complet.

Dans ce chapitre, nous allons détailler les équations du EM avec un Forward-Backward sur toutes les chaînes simultanément pour le sous-graphe du MHMM-DF associé à la dynamique des plantes annuelles. Par la suite, nous allons modifier les équations du EM avec un Forward-Backward chaîne par chaîne pour le rendre applicable au MHMM-DF complet. Le choix de paramétrisation des lois du modèles ayant un impact sur l'étape M du EM, les calculs seront explicités selon deux choix de paramétrisation. La première paramétrisation considère que toutes les lois du MHMM-DF suivent des lois Binomiales logistiques et la deuxième considère que toutes les lois du MHMM-DF sont des Binomiales logistiques, sauf pour la loi d'émission qui suit une loi Binomiale logistique avec états cachés non-éteints. On explicite aussi l'étape M avec des paramètres dépendant de la saison de culture du champ dans l'annexe.

Au-delà de l'estimation du modèle, nous allons également montrer comment efficacement calculer la vraisemblance des données observées, restaurer les états des variables cachées et prédire les états suivants des variables observées.

5.1 EM global sur le sous-graphe du MHMM-DF associé aux adventices

La démonstration que $\arg\max_\lambda \mathbb{P}(Y^{C,N}|\lambda)$ est un point fixe de l'algorithme EM pour le MHMM-DF est identique à celle donnée au chapitre 2.

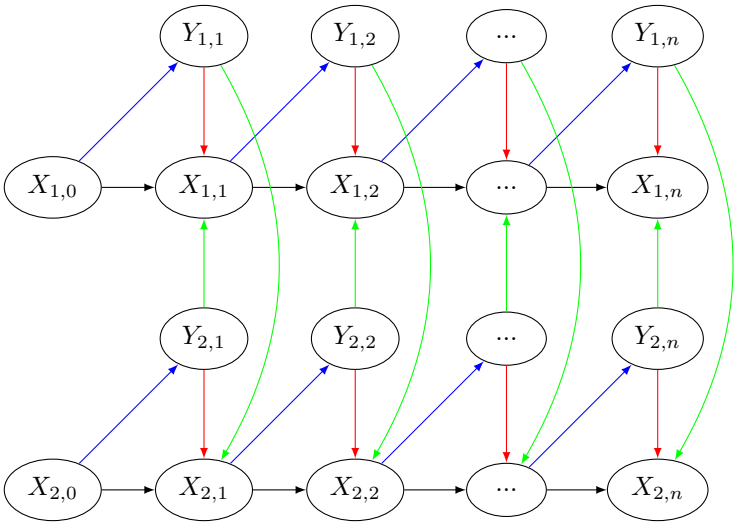


FIGURE 5.1 – Le sous graphe d'un MHMM-DF associé à la dynamique des plantes annuelles

L'expression de l'espérance conditionnelle du logarithme de la vraisemblance complète utilisée pour la mise à jour des paramètres est la suivante.

$$\begin{aligned}
E(X^{C,N}, Y^{C,N}, \lambda, \lambda_{it}) &= \mathbb{E}[\ln(\mathbb{P}(Y^{C,N}, X^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
&= \mathbb{E}[\ln(\prod_{c=1}^{C} \mathbb{P}(X_{c,0}|\lambda)|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
&\quad + \mathbb{E}[\ln(\prod_{c=1}^{C}\prod_{n=1}^{N} \mathbb{P}(X_{c,n}|X_{c,n-1}, Y_n^C, \lambda)\mathbb{P}(Y_{c,n}|X_{c,n-1}, \lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
&= \sum_{x^C \in \Omega_X^C} \ln(\prod_{c=1}^{C} \pi(x_c))\mathbb{P}(X_0^C = x^C|Y^{C,N} = y^{C,N}, \lambda_{it}) \\
&\quad + \sum_{n=1}^{N} \sum_{(x_n^C, x_{n-1}^C) \in \Omega_X^{2C}} \ln(\prod_{c=1}^{C} A(x_{c,n-1}, x_{c,n}, y_n^C)) \\
&\quad \times \mathbb{P}(X_n^C = x_n^C, X_{n-1}^C = x_{n-1}^C|Y^{C,N} = y^{C,N}, \lambda_{it}) \\
&\quad + \sum_{n=1}^{N} \sum_{x_{n-1}^C \in \Omega_X^C} \ln(\prod_{c=1}^{C} \phi(x_{c,n-1}, y_{c,n}))\mathbb{P}(X_{n-1}^C = x_{n-1}^C|Y^{C,N} = y^{C,N}, \lambda_{it})
\end{aligned}$$

On peut remarquer que dans l'expression de l'espérance conditionnelle du logarithme de la vraisemblance, certaines probabilités sont inconnues. On utilise l'étape E pour les calculer.

5.1.1 Étape E

Nous présentons ici l'application directe des formules du Forward-Backward global. Ce dernier est utilisé pour calculer les probabilités suivantes.

— $\rho_n(x_n^C) = \mathbb{P}(X_n^C = x_n^C|Y^{C,N} = y^{C,N}, \lambda_{it})$
— $\xi_n(x_{n-1}^C, x_n^C) = \mathbb{P}(X_{n-1}^C = x_{n-1}^C, X_n^C = x_n^C|Y^{C,N} = y^{C,N}, \lambda_{it})$

Pour ce faire, nous utilisons les fonctions α_n et β_n définies ci-dessous.

$$\alpha_n(x_n^C) = \mathbb{P}(Y^{C,n} = y^{C,n}, X_n^C = x_n^C|\lambda_{it})$$

$$\beta_n(x_n^C) = \mathbb{P}(Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_n^C|X_n^C = x_n^C, \lambda_{it})$$

Pour tout $n \in \{1, \ldots, N\}$,

$$\begin{aligned}
\alpha_n(x_n^C) &= \mathbb{P}(Y^{C,n} = y^{C,n}, X_n^C = x_n^C | \lambda_{it}) \\
&= \sum_{x_{n-1}^C \in \Omega_X^C} \mathbb{P}(Y^{C,n} = y^{C,n}, X_n^C = x_n^C, X_{n-1}^C = x_{n-1}^C | \lambda_{it}) \\
&= \sum_{x_{n-1}^C \in \Omega_X^C} (\mathbb{P}(Y^{C,n-1} = y^{C,n-1}, X_{n-1}^C = x_{n-1}^C | \lambda_{it}) \\
&\quad \times \mathbb{P}(Y_n^C = y_n^C, X_n^C = x_n^C | X_{n-1}^C = x_{n-1}^C, Y^{C,n-1} = y^{C,n-1}, \lambda_{it})\big) \\
&= \sum_{x_{n-1}^C \in \Omega_X^C} \big(\alpha_{n-1}(x_{n-1}^C) \mathbb{P}(X_n^C = x_n^C | Y_n^C = y_n^C, X_{n-1}^C = x_{n-1}^C, \lambda_{it}) \\
&\quad \times \mathbb{P}(Y_n^C = y_n^C | X_{n-1}^C = x_{n-1}^C, Y_{n-1}^C = y_{n-1}^C, \lambda_{it})\big) \\
&= \sum_{x_{n-1}^C \in \Omega_X^C} \Bigg(\alpha_{n-1}(x_{n-1}^C) \prod_{c=1}^{C} \big[\mathbb{P}(X_{c,n} = x_{c,n} | Y_n^C = y_n^C, X_{c,n-1} = x_{c,n-1}, \lambda_{it}) \\
&\quad \times \mathbb{P}(Y_{c,n} = y_{c,n} | X_{c,n-1} = x_{c,n-1}, Y_{n-1}^C = y_{n-1}^C, \lambda_{it})\big]\Bigg)
\end{aligned}$$

On obtient donc une relation de récurrence sur la suite (α_n) : pour tout $n \in \{1, \ldots, N\}$,

$$\alpha_n(x_n^C) = \sum_{x_{n-1}^C \in \Omega_X^C} \alpha_{n-1}(x_{n-1}^C) \prod_{c=1}^{C} A_{it}(x_{c,n-1}, x_{c,n}, y_n^C) \phi_{it}(x_{c,n-1}, y_{c,n})$$

avec $\alpha_0(x_{\mathbf{0}}^C) = \prod_{c=1}^{C} \pi_{it}(x_{c,0})$. De façon analogue, on détermine la relation de récurrence suivante pour β_n : pour tout $n \in \{1, \ldots, N\}$,

$$\begin{aligned}
\beta_n(x_n^C) &= \mathbb{P}(Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_n^C | X_n^C = x_n^C, \lambda_{it}) \\
&= \mathbb{P}(Y_{n+1}^C = y_{n+1}^C | X_n^C = x_n^C, \lambda_{it}) \\
&\quad \times \mathbb{P}(Y_{n+2}^C = y_{n+2}^C, ..., Y_N^C = y_n^C | X_n^C = x_n^C, Y_{n+1}^C = y_{n+1}^C, \lambda_{it}) \\
&= \sum_{x_{n+1}^C \in \Omega_X^C} \big(\mathbb{P}(Y_{n+2}^C = y_{n+2}^C, ..., Y_N^C = y_n^C, X_{n+1}^C = x_{n+1}^C | X_n^C = x_n^C, Y_{n+1}^C = y_{n+1}^C, \lambda_{it})\big) \\
&\quad \times \prod_{c=1}^{C} \phi_{it}(x_{c,n}, y_{c,n+1}) \\
&= \sum_{x_{n+1}^C \in \Omega_X^C} \big(\mathbb{P}(Y_{n+2}^C = y_{n+2}^C, ..., Y_N^C = y_n^C | X_{n+1}^C = x_{n+1}^C, X_n^C = x_n^C, Y_{n+1}^C = y_{n+1}^C, \lambda_{it}) \\
&\quad \times \mathbb{P}(X_{n+1}^C = x_{n+1}^C | X_n^C = x_n^C, Y_{n+1}^C = y_{n+1}^C, \lambda_{it})\big) \prod_{c=1}^{C} \phi_{it}(x_{c,n}, y_{c,n+1}) \\
&= \sum_{x_{n+1}^C \in \Omega_X^C} \mathbb{P}(Y_{n+2}^C = y_{n+2}^C, ..., Y_N^C = y_n^C | X_{n+1}^C = x_{n+1}^C, \lambda_{it}) \\
&\quad \times \prod_{c=1}^{C} \phi_{it}(x_{c,n}, y_{c,n+1}) A_{it}(x_{c,n}, x_{c,n+1}, y_{n+1}^C) \\
&= \sum_{x_{n+1}^C \in \Omega_X^C} \beta_n(x_{n+1}^C) \prod_{c=1}^{C} \phi_{it}(x_{c,n}, y_{c,n+1}) A_{it}(x_{c,n}, x_{c,n+1}, y_{n+1}^C)
\end{aligned}$$

avec $\beta_N(x_N^C) = 1$.

La relation suivante entre α_n et β_n est cruciale afin de pouvoir calculer les ρ et ξ.

$$\begin{aligned}
\alpha_n(x_n^C) \times \beta_n(x_n^C) = \quad & \mathbb{P}(Y^{C,n} = y^{C,n}, X_n^C = x_n^C | \lambda_{it}) \times \mathbb{P}(Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_n^C | X_n^C = x_n^C, \lambda_{it}) \\
= \quad & \mathbb{P}(X_n^C = x_n^C | \lambda_{it}) \times \mathbb{P}(Y^{C,n} = y^{C,n} | X_n^C = x_n^C, \lambda_{it}) \\
& \times \mathbb{P}(Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_n^C | X_n^C = x_n^C, \lambda_{it}) \qquad (5.1) \\
= \quad & \mathbb{P}(X_n^C = x_n^C | \lambda_{it}) \times \mathbb{P}(Y^{C,N} = y^{C,N} | X_n^C = x_n^C, \lambda_{it}) \qquad (5.2) \\
= \quad & \mathbb{P}(Y^{C,N} = y^{C,N}, X_n^C = x_n^C | \lambda_{it})
\end{aligned}$$

Le passage de l'équation (5.1) à (5.2) nécessite l'indépendance des $Y^{C,n}$ avec les $Y_{n+1}^C, \ldots, Y_N^C$ sachant X_n^C. Cette condition est vérifiée pour le sous graphe du MHMM-DF associé au plantes annuelles mais n'est pas vérifiée pour le MHMM-DF complet.

Pour appliquer le Forward-Backward chaîne par chaîne, il faudrait montrer que

$$\begin{aligned}
\mathbb{P}(Y^{C,N} = y^{C,N} | X_{c,n} = x_{c,n}, \lambda_{it}) = & \mathbb{P}(Y^{C,n} = y^{C,n} | X_{c,n} = x_{c,n}, \lambda_{it}) \\
& \times \mathbb{P}(Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_n^C | X_{c,n} = x_{c,n}, \lambda_{it}),
\end{aligned}$$

c'est-à-dire qu'il y a indépendance des $Y^{C,n}$ avec les $Y_{n+1}^C, \ldots, Y_N^C$ sachant la variable cachée de la chaîne c au temps n ($X_{c,n}$). Une telle hypothèse permettrait une complexité algorithmique linéaire. Cependant cette indépendance n'est vérifiée ni pour le MHMM-DF complet ni pour le sous-graphe des plantes annuelles.

On en déduit donc la vraisemblance des données à l'aide de (α_n) et (β_n).

$$\mathbb{P}(Y^{C,N} = y^{C,N} | \lambda_{it}) = \sum_{x_n^C \in \Omega_X^C} \alpha_n(x_n^C) \times \beta_n(x_n^C)$$

Ensuite, calculons $\rho_n(x_n^C)$ et $\xi_n(x_{n-1}^C, x_n^C)$ pour tout $(x_{n-1}^C, x_n^C) \in \Omega_X^C \times \Omega_X^C$.

$$\rho_n(x_n^C) = \mathbb{P}(X_n^C = x_n^C | Y^{C,N} = y^{C,N}, \lambda_{it}) = \frac{\alpha_n(x_n^C) \times \beta_n(x_n^C)}{\sum_{x_n^C \in \Omega_X^C} [\alpha_n(x_n^C) \times \beta_n(x_n^C)]}$$

$$\begin{aligned}
\xi_n(x_{n-1}^C, x_n^C) = & \mathbb{P}(X_{n-1}^C = x_{n-1}^C, X_n^C = x_n^C | Y^{C,N} = y^{C,N}, \lambda_{it}) \\
= & \frac{\mathbb{P}(X_n^C = x_n^C, X_{n-1}^C = x_{n-1}^C, Y_n^C = y_n^C | \lambda_{it})}{\mathbb{P}(Y^{C,N} = y^{C,N} | \lambda_{it})} \\
& \times \mathbb{P}(Y^{C,n-1} = y^{C,n-1}, Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_N^C | X_n^C = x_n^C, X_{n-1}^C = x_{n-1}^C, \lambda_{it}) \\
= & \frac{\mathbb{P}(X_n^C = x_n^C | X_{n-1}^C = x_{n-1}^C, Y_n^C = y_n^C, \lambda_{it}) \mathbb{P}(X_{n-1}^C = x_{n-1}^C, Y_n^C = y_n^C | \lambda_{it})}{\mathbb{P}(Y^{C,N} = y^{C,N} | \lambda_{it})} \\
& \times \mathbb{P}(Y^{C,n-1} = y^{C,n-1} | X_{n-1}^C = x_{n-1}^C, \lambda_{it}) \mathbb{P}(Y_{n+1}^C = y_{n+1}^C, ..., Y_N^C = y_N^C | X_n^C = x_n^C, \lambda_{it}) \\
= & \frac{\prod_{c=1}^{C} \left[A_{it}(x_{c,n-1}, x_{c,n}, y_n^C) \phi_{it}(x_{c,n-1}, y_{c,n}) \right] \mathbb{P}(X_{n-1}^C = x_{n-1}^C | \lambda_{it})}{\mathbb{P}(Y^{C,N} = y^{C,N} | \lambda_{it})} \\
& \times \mathbb{P}(Y^{C,n-1} = y^{C,n-1} | X_{n-1}^C = x_{n-1}^C, \lambda_{it}) \beta_n(x_n^C) \\
= & \frac{\prod_{c=1}^{C} \left[A_{it}(x_{c,n-1}, x_{c,n}, y_n^C) \phi_{it}(x_{c,n-1}, y_{c,n}) \right] \alpha_{n-1}(x_{n-1}^C) \beta_n(x_n^C)}{\mathbb{P}(Y^{C,N} = y^{C,N} | \lambda_{it})}
\end{aligned}$$

Le calcul de tous les α_n et β_n a une complexité algorithmique de $O(N|\Omega_X|^C)$. Il en est de même pour les ρ_n. En revanche, le calcul des ξ_n est en $O(N|\Omega_X|^{2C})$. Par conséquent, la complexité algorithmique de l'étape E avec le Forward-Backward global est de $O(N|\Omega_X|^{2C})$.

5.1.2 Étape M

Les paramètres étant des probabilités, on doit imposer des contraintes de normalité qui sont prises en compte par la méthode des multiplicateurs de Lagrange. Pour déterminer l'estimateur λ_{it+1} de $\lambda = (\pi, A, \phi)$, on estime indépendamment π, A et ϕ car la dérivée de la fonction de Lagrange par rapport à chacun d'entre eux ne dépend pas des deux autres. On pose

$$\eta_2 = (\eta_{2,1,1}, ..., \eta_{2,1,|\Omega_Y|^C}, \eta_{2,2,1}, ..., \eta_{2,|\Omega_X|,|\Omega_Y|^C})$$

et $\eta_3 = (\eta_{3,1}, ..., \eta_{3,|\Omega_X|})$. La fonction que l'on doit maximiser est la fonction L définie par

$$\begin{aligned} L(\pi, A, \eta_1, \eta_2, \eta_3) =& \mathbb{E}[\ln(\mathbb{P}(X^{C,N}, Y^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] - \eta_1 \left[\left(\sum_{x_0 \in \Omega_X} \pi(x_0) \right) - 1 \right] \\ & - \sum_{(x_{n-1}, y^C) \in \Omega_X \times \Omega_Y^C} \eta_{2,x_{n-1},y^C} \left[\left(\sum_{x_n \in \Omega_X} A(x_{n-1}, x_n, y^C) \right) - 1 \right] \\ & - \sum_{x \in \Omega_X} \eta_{3,x} \left[\left(\sum_{y \in \Omega_Y} \phi(x, y) \right) - 1 \right] \end{aligned}$$

Les dérivées partielles sont

$$\begin{aligned} \frac{\partial L(\pi, A, \phi, \eta_1, \eta_2, \eta_3)}{\partial \pi(x_0)} &= \sum_{\substack{j \in \Omega_X^C \\ j_c = x_0}} \frac{\rho_0(j)}{\pi(x_0)} - \eta_1 \\ \frac{\partial L(\pi, A, \phi, \eta_1, \eta_2, \eta_3)}{\partial \phi(x, y)} &= \sum_{\substack{(c,n) \\ y_c = y_{c,n}}} \sum_{\substack{j \in \Omega_X^C \\ j_c = x}} \frac{\rho_n(x)}{\phi(x, y)} - \eta_{3,x} \\ \frac{\partial L(\pi, A, \phi, \eta_1, \eta_2, \eta_3)}{\partial A(x, x', y^C)} &= \sum_{\substack{(c,n) \\ y^{\mathcal{C}\setminus c} = y_n^{\mathcal{C}\setminus c} \text{ et } y_c = y_{c,n}}} \sum_{\substack{(j_{n-1}, j_n) \in (\Omega_X^C)^2 \\ j_{c,n-1} = x \text{ et } j_{c,n} = x'}} \frac{\xi_n(j_{n-1}, j_n)}{A(x, x', y^C)} - \eta_{2,x,y^C} \\ \frac{\partial L(\pi, A, \phi, \eta_1, \eta_2, \eta_3)}{\partial \eta_1} &= \left(\sum_{x_0 \in \Omega_X} \pi(x_0) \right) - 1 \\ \frac{\partial L(\pi, A, \phi, \eta_1, \eta_2, \eta_3)}{\partial \eta_{2,x,y^C}} &= \left(\sum_{x_n \in \Omega_X} A(x, x_n, y^C) \right) - 1 \\ \frac{\partial L(\pi, A, \phi, \eta_1, \eta_2, \eta_3)}{\partial \eta_{3,x}} &= \left(\sum_{y \in \Omega_Y} \phi(x, y) \right) - 1 \end{aligned}$$

On obtient donc

$$\pi_{it+1}(x_0) = \sum_{\substack{j\in\Omega_X^C \\ j_c=x_0}} \rho_0(j)$$

$$\phi_{it+1}(x,y) = \frac{\sum_{\substack{(c,n) \\ y=y_{c,n}}} \sum_{\substack{j\in\Omega_X^C \\ j_c=x}} \rho_n(j)}{\sum_{\substack{(c,n) \\ y=y_{c,n}}} \sum_{j'\in\Omega_X^C} \rho_n(j')}$$

$$A_{it+1}(x,x',y^C) = \frac{\sum_{\substack{(c,n) \\ y^{\mathcal{C}\setminus c}=y_n^{\mathcal{C}\setminus c} \text{ et } y_c=y_{c,n}}} \sum_{\substack{(j_{n-1},j_n)\in(\Omega_X^C)^2 \\ j_{c,n-1}=x \text{ et } j_{c,n}=x'}} \xi_n(j_{n-1},j_n)}{\sum_{\substack{(c,n) \\ y^{\mathcal{C}\setminus c}=y_n^{\mathcal{C}\setminus c} \text{ et } y_c=y_{c,n}}} \sum_{\substack{(i_{n-1},i_n)\in(\Omega_X^C)^2 \\ i_{c,n-1}=x}} \xi_n(i_{n-1},i_n)}$$

La complexité algorithmique de l'étape M dépend de la complexité algorithmique de la mise à jour de A_{it+1}. Le calcul de tous les ϕ_{it+1} et les π_{it+1} a une complexité algorithmique de $O(N|\Omega_X|^C)$. Pour calculer les coefficients de A_{it+1}, il faut faire une somme sur tous les pas de temps et deux sommes sur toutes les variables cachées de départs et d'arrivées possibles. La complexité algorithmique du calcul de tous les coefficients de A_{it+1}, et donc de l'étape M, est en $O(N|\Omega_X|^C)$.

5.2 EM pour MHHM-DF complet

Dans cette section, nous allons montrer comment implémenter de manière plus efficace l'algorithme EM sur le MHMM-DF complet. Ici, le Forward-Backward est implémenté chaîne par chaîne tout en conservant l'exactitude de l'étape E.

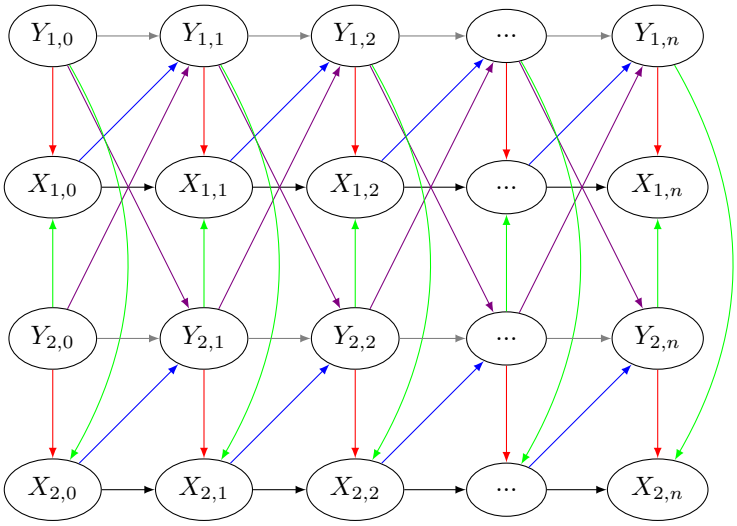


FIGURE 5.2 – Le graphe du MHMM-DF complet

L'expression de l'espérance du logarithme de la vraisemblance du modèle peut être détaillée

chaîne par chaîne.

$$
\begin{aligned}
E(X^{C,N}, Y^{C,N}, \lambda, \lambda_{it}) =& E[\ln(\mathbb{P}(Y^{C,N}, X^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
=& E[\ln(\prod_{c=1}^{C} \mathbb{P}(X_{c,0}|Y_{c,0}, \lambda)\mathbb{P}(Y_{c,0}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
&+ E[\ln(\prod_{c=1}^{C}\prod_{n=1}^{N} \mathbb{P}(X_{c,n}|X_{c,n-1}, Y_n^C, \lambda)\mathbb{P}(Y_{c,n}|X_{c,n-1}, Y_{n-1}^C, \lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
=& \sum_{c=1}^{C} \ln(\zeta(y_{c,0})) \\
&+ \sum_{c=1}^{C}\sum_{x\in\Omega_X} \ln(\pi(x))\mathbb{P}(X_{c,0} = x|Y^{C,N} = y^{C,N}, \lambda_{it}) \\
&+ \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{(x_n,x_{n-1})\in\Omega_X^2} \ln(A(x_{n-1}, x_n, y_n^C))\mathbb{P}(x_n, x_{n-1}|y^{C,N}, \lambda_{it}) \\
&+ \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{n-1}\in\Omega_X} \ln(\phi(x_{n-1}, y_{c,n}, y_{n-1}^C))\mathbb{P}(X_{c,n-1} = x_{n-1}|Y^{C,N} = y^{C,N}, \lambda_{it}) \\
=& \sum_{c=1}^{C} \ln(\zeta(y_{c,0})) \\
&+ \sum_{c=1}^{C}\sum_{x\in\Omega_X} \ln(\pi(x))\rho_{c,0}(x) \\
&+ \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{(x_n,x_{n-1})\in\Omega_X^2} \ln(A(x_{n-1}, x_n, y_n^C))\xi_{c,n}(x_{n-1}, x_n) \\
&+ \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{n-1}\in\Omega_X} \ln(\phi(x_{n-1}, y_{c,n}, y_{n-1}^C))\rho_{c,n-1}(x_{n-1})
\end{aligned}
$$

5.2.1 Étape E

Le but de l'étape E est de calculer $\rho_{c,n}(x_{c,n})$ et $\xi_{c,n}(x_{c,n-1}, x_{c,n})$ définies pour tous $n \in \{1, \ldots, N\}$ et $c \in \{1, \ldots, C\}$ par

$$
\begin{aligned}
\xi_{c,n}(x_{n-1}, x_n) =& \mathbb{P}(X_{c,n} = x_n, X_{c,n-1} = x_{n-1}|Y^{C,N} = y^{C,N}, \lambda_{it}) \\
\rho_{c,n}(x_n) =& \mathbb{P}(X_{c,n} = x_n|Y^{C,N} = y^{C,N}, \lambda_{it})
\end{aligned}
$$

où $\lambda_{it} = (\zeta_{it}, \pi_{it}, \phi_{it}, A_{it})$ est la version courante des estimateurs.

Pour calculer ξ et ρ, on doit modifier $\beta_{c,n}$ pour conserver la propriété suivante.

$$\alpha_{c,n}(x_{c,n}) \times \beta_{c,n}(x_{c,n}) = \mathbb{P}(y^{C,N}, X_{c,n} = x_{c,n}|\lambda_{it})$$

On pose donc

$$
\begin{aligned}
\alpha_{c,n}(x_{c,n}) &= \mathbb{P}(y^{C,n}, X_{c,n} = x_{c,n}|\lambda_{it}) \\
\beta_{c,n}(x_{c,n}) &= \mathbb{P}(y_{n+1}^C, ..., y_N^C|y^{C,n}, x_{c,n}, \lambda_{it})
\end{aligned}
$$

Pour un patch c donné, $\alpha_{c,n}$ peut être calculée de façon récursive.

$$\begin{aligned}
\alpha_{c,n}(x_{c,n}) &= \mathbb{P}(y^{C,n}, X_{c,n} = x_{c,n}|\lambda_{it}) \\
&= \sum_{x_{c,n-1}\in\Omega_X} \mathbb{P}(y^{C,n}, X_{c,n} = x_{c,n}, X_{c,n-1} = x_{c,n-1}|\lambda_{it}) \\
&= \sum_{x_{c,n-1}\in\Omega_X} \mathbb{P}(y^{C,n-1}, X_{c,n-1} = x_{c,n-1}|\lambda_{it})\mathbb{P}(y_n^C, X_{c,n} = x_{c,n}|y^{C,n-1}, x_{c,n-1}, \lambda_{it}) \\
&= \sum_{x_{c,n-1}\in\Omega_X} \alpha_{c,n-1}(x_{c,n-1})\mathbb{P}(y_n^C|y^{C,n-1}, x_{c,n-1}, \lambda_{it})A_{it}(x_{c,n-1}, x_{c,n}, y_n^C) \\
&= \sum_{x_{c,n-1}\in\Omega_X} \alpha_{c,n-1}(x_{c,n-1})\phi_{it}(x_{c,n-1}, y_{c,n}, y_{n-1}^C) \\
&\quad \times\mathbb{P}(y_n^{\mathcal{C}\setminus c}|y^{C,n-1}, y_{c,n}, x_{c,n-1}, \lambda_{it})A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)
\end{aligned}$$

Puisque $Y_n^{\mathcal{C}\setminus c}$ est indépendant de $X_{c,n-1}$ et de $Y_{c,n}$, conditionnellement à $Y^{C,n-1}$, on en déduit que $\mathbb{P}(y_n^{\mathcal{C}\setminus c}|y^{C,n-1}, y_{c,n}, x_{c,n-1}) = \mathbb{P}(y_n^{\mathcal{C}\setminus c}|y^{C,n-1})$ est une constante que l'on notera $K_{c,n}$. Par conséquent

$$\alpha_{c,n}(x_{c,n}) \propto \sum_{x_{c,n-1}\in\Omega_X} \alpha_{c,n-1}(x_{c,n-1})\phi_{it}(x_{c,n-1}, y_{c,n}, y_n^C)A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)$$

En pratique, à la place de calculer $\alpha_{c,n}(x_{c,n})$, on calcule $\tilde{\alpha}_{c,n}(x_{c,n})$ définie par :

$$\tilde{\alpha}_{c,n}(x_{c,n}) = \sum_{x_{c,n-1}\in\Omega_X} \tilde{\alpha}_{c,n-1}(x_{c,n-1})\phi_{it}(x_{c,n-1}, y_{c,n}, y_n^C)A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)$$

avec $\tilde{\alpha}_{c,0}(x_{c,0}) = \alpha_{c,0}(x_{c,0})$

De façon similaire, $\beta_{c,n}$ peut être calculée de façon récursive.

$$\begin{aligned}
\beta_{c,n}(x_{c,n}) =& \mathbb{P}(y_{n+1}^C, ..., y_N^C|y^{C,n}, x_{c,n}, \lambda_{it}) \\
&= \sum_{x_{c,n+1}\in\Omega_X} \mathbb{P}(y_{n+1}^C, ..., y_N^C, X_{c,n+1} = x_{c,n+1}|y^{C,n}, x_{c,n}, \lambda_{it}) \\
&= \sum_{x_{c,n+1}\in\Omega_X} \mathbb{P}(y_{n+2}^C, ..., y_N^C, X_{c,n+1} = x_{c,n+1}|y^{C,n+1}, x_{c,n}, \lambda_{it})\mathbb{P}(y_{n+1}^C|y^{C,n}, x_{c,n}, \lambda_{it}) \\
&= \sum_{x_{c,n+1}\in\Omega_X} \mathbb{P}(y_{n+2}^C, ..., y_N^C|y^{C,n+1}, x_{c,n}, x_{c,n+1}, \lambda_{it})A_{it}(x_{c,n}, x_{c,n+1}, y_{n+1}^C) \\
&\quad \times \phi_{it}(x_{c,n}, y_{c,n+1}, y_n^C)\mathbb{P}(y_{n+1}^{\mathcal{C}\setminus c}|y^{C,n}, y_{c,n+1}, x_{c,n}, \lambda_{it}) \\
=& K_{c,n+1} \sum_{x_{c,n+1}\in\Omega_X} \mathbb{P}(y_{n+2}^C, ..., Y_N^C|y^{C,n+1}, x_{c,n}, x_{c,n+1}, \lambda_{it}) \\
&\quad \times A_{it}(x_{c,n}, x_{c,n+1}, y_{n+1}^C)\phi_{it}(x_{c,n}, y_{c,n+1}, y_n^C)
\end{aligned}$$

Donc

$$\beta_{c,n}(x_{c,n}) \propto \sum_{x_{c,n+1}\in\Omega_X} \beta_{c,n+1}(x_{c,n+1})A_{it}(x_{c,n}, x_{c,n+1}, Y_{n+1}^C)\phi_{it}(x_{c,n}, y_{c,n+1}, y_n^C).$$

On considère $\tilde{\beta}_{c,n}(x_{c,n})$ définie par :

$$\tilde{\beta}_{c,n}(x_{c,n}) = \sum_{x_{c,n+1}\in\Omega_X} \tilde{\beta}_{c,n+1}(x_{c,n+1})A_{it}(x_{c,n}, x_{c,n+1}, y_{n+1}^C)\phi_{it}(x_{c,n}, y_{c,n+1}, y_n^C)$$

où $\tilde{\beta}_{c,N}(x_{c,N}) = \beta_{c,N}(x_{c,N})$.

Calculons maintenant $\rho_{c,n}(x_{c,n})$ et $\xi_{c,n}(x_{c,n-1}, x_{c,n})$ à l'aide de $\tilde{\alpha}_{c,n}$ et $\tilde{\beta}_{c,n}$.

$$\rho_{c,n}(x_{c,n}) = \mathbb{P}(X_{c,n} = x_{c,n}|y^{C,N}, \lambda_{it}) = \frac{\mathbb{P}(X_{c,n} = x_{c,n}, y^{C,N}|\lambda_{it})}{\mathbb{P}(y^{C,N}|\lambda_{it})}$$
$$= \frac{\beta_{c,n}(x_{c,n})\alpha_{c,n}(x_{c,n})}{\sum_{x\in\Omega_X}\beta_{c,n}(x)\alpha_{c,n}(x)} = \frac{\tilde{\beta}_{c,n}(x_{c,n})\tilde{\alpha}_{c,n}(x_{c,n})}{\sum_{x\in\Omega_X}\tilde{\beta}_{c,n}(x)\tilde{\alpha}_{c,n}(x)}$$

$$\xi_{c,n}(x_{c,n-1}, x_{c,n}) = \mathbb{P}(X_{c,n} = x_{c,n}, X_{c,n-1} = x_{c,n-1}|y^{C,N}, \lambda_{it}) = \frac{\mathbb{P}(X_{c,n} = x_{c,n}, X_{c,n-1} = x_{c,n-1}, y^{C,N}|\lambda_{it})}{\mathbb{P}(y^{C,N}|\lambda_{it}))}$$

Il est possible d'exprimer le numérateur en termes de $\tilde{\alpha}_{c,n}$ et $\tilde{\beta}_{c,n}$

$$\begin{aligned}
\mathbb{P}(x_{c,n}, x_{c,n-1}, y^{C,N}) &= \mathbb{P}(X_{c,n} = x_{c,n}, X_{c,n-1} = x_{c,n-1}, y_n^C|\lambda_{it}) \\
&\quad \times\mathbb{P}(y^{C,N\setminus n}|x_{c,n}, x_{c,n-1}, y_n^C, \lambda_{it}) \\
&= A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\mathbb{P}(X_{c,n-1} = x_{c,n-1}, y_n^C|\lambda_{it}) \\
&\quad \times\mathbb{P}(y^{C,n-1}|x_{c,n}, x_{c,n-1}, y_n^C, \lambda_{it}) \\
&\quad \times\mathbb{P}(y_{n+1}^C, ..., y_N^C|x_{c,n}, x_{c,n-1}, y^{C,n}, \lambda_{it}) \\
&= A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\mathbb{P}(X_{c,n-1} = x_{c,n-1}, y_n^C|\lambda_{it}) \\
&\quad \times\mathbb{P}(y^{C,n-1}|x_{c,n-1}, y_n^C, \lambda_{it})\beta_{c,n}(x_{c,n}) \\
&= A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\beta_{c,n}(x_{c,n})\mathbb{P}(X_{c,n-1} = x_{c,n-1}, y^{C,n}|\lambda_{it}) \\
&= A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\beta_{c,n}(x_{c,n})\mathbb{P}(X_{c,n-1} = x_{c,n-1}, y^{C,n-1}|\lambda_{it}) \\
&\quad \times\mathbb{P}(y_n^C|x_{c,n-1}, y^{C,n-1}, \lambda_{it}) \\
&= A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\beta_{c,n}(x_{c,n})\alpha_{c,n-1}(x_{c,n-1})\phi_{it}(x_{c,n-1}, y_{c,n}, y_{n-1}^C) \\
&\propto A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\tilde{\beta}_{c,n}(x_{c,n})\tilde{\alpha}_{c,n-1}(x_{c,n-1})\phi_{it}(x_{c,n-1}, y_{c,n}, y_{n-1}^C)
\end{aligned}$$

Ainsi on obtient

$$\xi_{c,n}(x_{c,n-1}, x_{c,n}) = \frac{A_{it}(x_{c,n-1}, x_{c,n}, y_n^C)\tilde{\beta}_{c,n}(x_{c,n})\tilde{\alpha}_{c,n-1}(x_{c,n-1})\phi_{it}(x_{c,n-1}, y_{c,n}, y_n^C)}{\sum_{(x,x')\in\Omega_X^2} A_{it}(x', x, y_n^C)\tilde{\beta}_{c,n}(x)\tilde{\alpha}_{c,n-1}(x')\phi_{it}(x', y_{c,n}, y_n^C)}$$

Comme précédemment, le calcul de tous les $\xi_{c,n}$ détermine la complexité algorithmique de l'étape E, à savoir $O(CN|\Omega_X|^2)$.

5.2.2 Étape M

L'étape M consiste à mettre à jour les paramètres λ_{it} en maximisant l'espérance du logarithme de la vraisemblance du MHHM-DF complet conditionnellement aux données et à la valeur courante des paramètres.

Pour maximiser la fonction $E(X^{C,N}, Y^{C,N}, \lambda, \lambda_{it})$ on cherche le lieu d'annulation des dérivées partielles. L'étape M varie selon la modélisation des lois du modèle. Par conséquent, nous allons détailler les calculs de l'étape M pour plusieurs paramétrisations.

5.2.2.1 L'étape M avec des lois non paramétriques

Voici la valeur des paramètres qui annulent les dérivées partielles de la fonction $E(X^{C,N}, Y^{C,N}, \lambda, \lambda_{it})$.

$$\begin{aligned}
\zeta_{it+1}(y) =& \frac{1}{|\Omega_Y|} \\
\pi_{it+1}(x_0) =& \rho_0(x_0) \\
\phi_{it+1}(x,y) =& \frac{\sum\limits_{\substack{(c,n)\\ y=y_{c,n}}} \rho_n(j)}{\sum\limits_{\substack{(c,n)\\ y=y_{c,n}}} \sum\limits_{j' \in \Omega_X} \rho_n(j')} \\
A_{it+1}(x,x',y^C) =& \frac{\sum\limits_{\substack{(c,n)\\ y^{\mathcal{C}\setminus c}=y_n^{\mathcal{C}\setminus c} \text{ et } y_c=y_{c,n}}} \xi_n(x,x')}{\sum\limits_{\substack{(c,n)\\ y^{\mathcal{C}\setminus c}=y_n^{\mathcal{C}\setminus c} \text{ et } y_c=y_{c,n}}} \sum\limits_{i_n \in \Omega_X} \xi_n(x,i_n)}
\end{aligned}$$

La complexité de l'étape M est égale à celle de la mise à jour des coefficients de A_{it} et est en $O(CN|\Omega_X|^2)$.

5.2.2.2 L'étape M pour des lois Binomiales logistiques

Afin de pouvoir simplifier les équations de la dérivée de l'espérance du modèle, nous allons rappeler quelques notations.

$$\begin{aligned}
\phi(x_{n-1}, y_{c,n}, y_{n-1}^C) &= \binom{|\Omega_Y|-1}{y_{c,n}} \left[\frac{1}{1+\mathrm{e}^{-w_\mu}}\right]^{|\Omega_Y|-1} \left[\mathrm{e}^{-w_\mu}\right]^{|\Omega_Y|-y_{c,n}-1} \\
\pi(x,y) &= \binom{|\Omega_X|-1}{x} \left[\frac{1}{1+\mathrm{e}^{-z}}\right]^{|\Omega_X|-1} \left[\mathrm{e}^{-z}\right]^{|\Omega_X|-x-1} \\
\zeta(y) &= \binom{|\Omega_Y|-1}{y} \left[\frac{1}{1+\mathrm{e}^{-\theta}}\right]^{|\Omega_Y|-1} \left[\mathrm{e}^{-\theta}\right]^{|\Omega_Y|-y-1} \\
A(x_{n-1}, x_n, y_n^C) &= \binom{|\Omega_X|-1}{x_n} \left[\frac{1}{1+\mathrm{e}^{-w_\nu}}\right]^{|\Omega_X|-1} \left[\mathrm{e}^{-w_\nu}\right]^{|\Omega_X|-x_n-1}
\end{aligned}$$

où

$$\begin{aligned}
w_\nu &= \nu_1 \times \frac{x_{n-1}}{|\Omega_X|} + \nu_2 \times \frac{y_{c,n}}{|\Omega_Y|} + \nu_3 \times \frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|} + \nu_0 \\
w_\mu &= \mu_1 \times \frac{x_{n-1}}{|\Omega_X|} + \mu_2 \times \frac{y_{c,n}}{|\Omega_Y|} + \mu_3 \times \frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|} + \mu_0 \\
z &= \tau_0 + \tau_2 \frac{y_{c,0}}{|\Omega_Y|} + \tau_3 \frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|}
\end{aligned}$$

Détaillons la dérivée selon ν_2 du logarithme de A.

$$
\begin{aligned}
\frac{\partial \ln(A(x_{n-1},x_n,y_n^C))}{\partial \nu_2} =&(1-|\Omega_X|)\frac{-\frac{y_{c,n}}{|\Omega_Y|}\,\mathrm{e}^{-(\nu_1\times\frac{x_{n-1}}{|\Omega_X|}+\nu_2\times\frac{y_{c,n}}{|\Omega_Y|}+\nu_3\times\frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|}+\nu_0)}}{1+\mathrm{e}^{-(\nu_1\times\frac{x_{n-1}}{|\Omega_X|}+\nu_2\times\frac{y_{c,n}}{|\Omega_Y|}+\nu_3\times\frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|}+\nu_0)}}\\
&-(|\Omega_X|-x_n-1)(\frac{y_{c,n}}{|\Omega_Y|})\\
=&(|\Omega_X|-1)\frac{y_{c,n}}{|\Omega_Y|}+\frac{(1-|\Omega_X|)\frac{y_{c,n}}{|\Omega_Y|}}{1+\mathrm{e}^{-(\nu_1\times\frac{x_{n-1}}{|\Omega_X|}+\nu_2\times\frac{y_{c,n}}{|\Omega_Y|}+\nu_3\times\frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|}+\nu_0)}}\\
&-(|\Omega_X|-x_n-1)(\frac{y_{c,n}}{|\Omega_Y|})\\
=&(x_n)\frac{y_{c,n}}{|\Omega_Y|}+\frac{(1-|\Omega_X|)\frac{y_{c,n}}{|\Omega_Y|}}{1+\mathrm{e}^{-(\nu_1\times\frac{x_{n-1}}{|\Omega_X|}+\nu_2\times\frac{y_{c,n}}{|\Omega_Y|}+\nu_3\times\frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|}+\nu_0)}}
\end{aligned}
$$

Maintenant détaillons l'expression des dérivées partielles de la fonction $E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})$ par rapport aux paramètres du MHMM-DF.

$$
\begin{aligned}
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \theta} =&\sum_{c=1}^{C}\left[y_{c,0}+(1-|\Omega_Y|)\frac{1}{1+\mathrm{e}^{-\theta}}\right]\\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \tau_0} =&\sum_{c=1}^{C}\sum_{x_{c,0}\in\Omega_X}\left[(x_{c,0})+(1-|\Omega_X|)\frac{1}{1+\mathrm{e}^{-(\tau_0+\tau_2\frac{y_{c,0}}{|\Omega_Y|}+\tau_3\frac{f(y_0^{C-1})}{|f(\Omega_Y^{C-1})|})}}\right]\rho_{c,0}(x_{c,0})\\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \tau_2} =&\sum_{c=1}^{C}\sum_{x_{c,0}\in\Omega_X}\left[x_{c,0}\frac{y_{c,0}}{|\Omega_Y|}+(1-|\Omega_X|)\frac{y_{c,0}}{|\Omega_Y|}\frac{1}{1+\mathrm{e}^{-(\tau_0+\tau_2\frac{y_{c,0}}{|\Omega_Y|}+\tau_3\frac{f(y_0^{C-1})}{|f(\Omega_Y^{C-1})|})}}\right]\rho_{c,0}(x_{c,0})\\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \tau_3} =&\sum_{c=1}^{C}\sum_{x_{c,0}\in\Omega_X}\left[x_{c,0}\frac{f(y_0^{C-1})}{|f(\Omega_Y^{C-1})|}+(1-|\Omega_X|)\frac{f(y_0^{C-1})}{|f(\Omega_Y^{C-1})|}\frac{1}{1+\mathrm{e}^{-(\tau_0+\tau_2\frac{y_{c,0}}{|\Omega_Y|}+\tau_3\frac{f(y_0^{C-1})}{|f(\Omega_Y^{C-1})|})}}\right]\rho_{c,0}(x_{c,0})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_0} =&\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{(x_n,x_{n-1})\in\Omega_X^2}B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})\\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_1} =&\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{(x_n,x_{n-1})\in\Omega_X^2}\frac{x_{c,n-1}}{|\Omega_X|}B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})\\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_2} =&\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{(x_n,x_{n-1})\in\Omega_X^2}\frac{y_{c,n}}{|\Omega_Y|}B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})\\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_3} =&\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{(x_n,x_{n-1})\in\Omega_X^2}\frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|}B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})
\end{aligned}
$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_0} = \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_1} = \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X} \frac{x_{c,n-1}}{|\Omega_X|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_2} = \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X} \frac{y_{c,n-1}}{|\Omega_Y|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_3} = \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X} \frac{f(y_{n-1}^{C-1})}{|f(\Omega_Y^{C-1})|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

où $B_{c,n} = x_{c,n} + (1-|\Omega_X|)\frac{1}{1+\mathrm{e}^{-(\nu_0+\nu_1\frac{x_{c,n-1}}{|\Omega_X|}+\nu_2\frac{y_{c,n}}{|\Omega_Y|}+\nu_3\frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|})}}$ et

$B'_{c,n} = y_{c,n} + (1-|\Omega_Y|)\frac{1}{1+\mathrm{e}^{-(\mu_0+\mu_1\frac{x_{c,n-1}}{|\Omega_X|}+\mu_2\frac{y_{c,n-1}}{|\Omega_Y|}+\mu_3\frac{f(y_{n-1}^{C-1})}{|f(\Omega_Y^{C-1})|})}}$

La paramétrisation Binomiale logistique ne permet pas d'obtenir une expression explicite de la mise à jour des paramètres. Par conséquent, il est nécessaire d'utiliser une méthode de type descente de gradient pour mettre à jour les paramètres.

5.2.2.3 L'étape M pour des lois Binomiales logistiques et une Binomiale logistique uniquement pour les états cachés non-éteints

Dans ce modèle, seule la mise à jour pour des paramètres de μ est différente de la mise à jour exposée précédemment. Les dérivées partielles selon les paramètres μ_i sont

$$\begin{aligned}
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_0} &= \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X\setminus\{0\}} B'_{c,n}\rho_{c,n-1}(x_{c,n-1}) \\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_1} &= \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X\setminus\{0\}} \frac{x_{c,n-1}}{|\Omega_X|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1}) \\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_2} &= \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X\setminus\{0\}} \frac{y_{c,n-1}}{|\Omega_Y|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1}) \\
\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_3} &= \sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{x_{c,n-1}\in\Omega_X\setminus\{0\}} \frac{f(y_{n-1}^{C-1})}{|f(\Omega_Y^{C-1})|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})
\end{aligned}$$

5.3 Expression de la vraisemblance du MHMM-DF

A la différence du Forward-Backward pour HMM, le Forward-Backward modifié sur le MHMM-DF complet n'exige pas de calculer la vraisemblance des données. Cependant, celle-ci peut être utilisée afin d'identifier le modèle le plus adéquat pour un jeu de données. Nous allons montrer comment calculer la vraisemblance des données à l'aide des probabilités $\tilde{\alpha}_{c,n}$ et $\tilde{\beta}_{c,n}$.

$$
\begin{aligned}
\mathbb{P}(y^{C,N}) &= \sum_{x_{c,n}\in\Omega_X} \mathbb{P}(X_{c,n}=x_{c,n}, y^{C,N}) \\
&= \sum_{x_{c,n}\in\Omega_X} \alpha_{c,n}(x_{c,n})\beta_{c,n}(x_{c,n}) \\
&= \sum_{x_{c,n}\in\Omega_X} [\prod_{n''=1}^{n} K_{c,n''}]\tilde{\alpha}_{c,n}(x_{c,n})[\prod_{n'=n+1}^{N} K_{c,n'}]\tilde{\beta}_{c,n}(x_{c,n}) \\
&= [\prod_{n'=1}^{N} K_{c,n'}] \sum_{x_{c,n}\in\Omega_X} \tilde{\alpha}_{c,n}(x_{c,n})\tilde{\beta}_{c,n}(x_{c,n})
\end{aligned}
$$

Le calcul de la vraisemblance requiert de calculer $\prod_{n=1}^{N} K_{c,n}$. Cependant montrons d'abord que $\frac{\tilde{\alpha}_{c,n}(x_{c,n})}{\sum_{x\in\Omega_X}\tilde{\alpha}_{c,n}(x)} = \mathbb{P}(X_{c,n}=x_{c,n}|y^{C,n})$.

On sait que $\alpha_{c,n}(x) \propto \tilde{\alpha}_{c,n}(x)$ et que $\alpha_{c,n}(x) = \mathbb{P}(X_{c,n}=x_{c,n}, y^{C,n}) = \mathbb{P}(X_{c,n}=x_{c,n}|y^{C,n})\mathbb{P}(y^{C,n})$. Donc $\tilde{\alpha}_{c,n}(x) \propto \mathbb{P}(X_{c,n}=x_{c,n}|y^{C,n})$. Si l'on normalise, on obtient $\frac{\tilde{\alpha}_{c,n}(x_{c,n})}{\sum_{x\in\Omega_X}\tilde{\alpha}_{c,n}(x)} = \mathbb{P}(X_{c,n}=x_{c,n}|y^{C,n})$.

Calculons maintenant $\prod_{n=1}^{N} K_{c,n}$.

$$
\begin{aligned}
\prod_{n=1}^{N} K_{c,n} &= \prod_{n=1}^{N} \mathbb{P}(y_n^{\mathcal{C}\setminus c}|y^{C,n-1}) \\
&= \prod_{n=1}^{N}\prod_{l\in\{1,\dots,C\}\setminus\{c\}} \mathbb{P}(y_{l,n}|y^{C,n-1}) \\
&= \prod_{n=1}^{N}\prod_{l\in\{1,\dots,C\}\setminus\{c\}}\sum_{x_{l,n-1}\in\Omega_X} \mathbb{P}(y_{l,n}, X_{l,n-1}=x_{l,n-1}|y^{C,n-1}) \\
&= \prod_{n=1}^{N}\prod_{l\in\{1,\dots,C\}\setminus\{c\}}\sum_{x_{l,n-1}\in\Omega_X} \mathbb{P}(y_{l,n}|X_{l,n-1}=x_{l,n-1}, y^{C,n-1})\mathbb{P}(X_{l,n-1}=x_{l,n-1}|y^{C,n-1}) \\
\prod_{n=1}^{N} K_{c,n} &= \prod_{n=1}^{N}\prod_{l\in\{1,\dots,C\}\setminus\{c\}}\sum_{x_{l,n-1}\in\Omega_X} \phi(x_{l,n-1}, y_{l,n}, y_{n-1}^{C})\frac{\tilde{\alpha}_{l,n-1}(x_{l,n-1})}{\sum_{x\in\Omega_X}\tilde{\alpha}_{l,n-1}(x)}
\end{aligned}
$$

Finalement

$$
\mathbb{P}(y^{C,N}) = \left[\prod_{n'=1}^{N}\prod_{l\in\{1,\dots,C\}\setminus\{c\}}\sum_{x_{l,n'-1}\in\Omega_X} \phi(x_{l,n'-1}, y_{l,n'}, y_{n'-1}^{C})\frac{\tilde{\alpha}_{l,n'-1}(x_{l,n'-1})}{\sum_{x\in\Omega_X}\tilde{\alpha}_{l,n'-1}(x)}\right]\left[\sum_{x_{c,n}\in\Omega_X}\tilde{\alpha}_{c,n}(x_{c,n})\tilde{\beta}_{c,n}(x_{c,n})\right]
$$

La complexité algorithmique de la vraisemblance est en $O(CN|\Omega_X|)$.

5.4 Viterbi pour MHMM-DF

L'algorithme de Viterbi (Forney, 1973) permet dans un HMM de récupérer la trajectoire d'états cachés la plus probable d'après les observations. De la même manière que pour le Forward-Backward, l'application telle quelle de cet algorithme au MHMM-DF complet a une complexité algorithmique exponentielle en C. Nous montrons comment réduire cette complexité en mettant

en oeuvre l'algorithme chaîne par chaîne. Ceci est possible car on peut décomposer la probabilité jointe des variables observées et cachées comme produit des probabilités des chaînes cachées sachant toutes les observations.

L'objectif du Viterbi est de calculer

$$\underset{x^{C,N}\in\Omega_X^{C,N}}{\arg\max}\ \mathbb{P}(X^{C,N}=x^{C,N},Y^{C,N}=y^{C,N})$$

Écrivons la probabilité jointe des variables cachées comme produit des probabilités des chaînes cachées sachant toutes les observations.

$$\begin{aligned}
\mathbb{P}(X^{C,N}=x^{C,N},Y^{C,N}=y^{C,N}) =& \mathbb{P}(X^{C,N}=x^{C,N}|Y^{C,N}=y^{C,N})\times\mathbb{P}(Y^{C,N}=y^{C,N})\\
=& \mathbb{P}(Y^{C,N}=y^{C,N})\prod_{c=1}^{C}p(X_c^N=x_c^N|Y^{C,N}=y^{C,N})\\
=& \mathbb{P}(Y^{C,N}=y^{C,N})\prod_{c=1}^{C}\frac{p(X_c^N=x_c^N,Y^{C,N}=y^{C,N})}{\mathbb{P}(Y^{C,N}=y^{C,N})}\\
=& \frac{1}{(\mathbb{P}(Y^{C,N}=y^{C,N}))^{C-1}}\prod_{c=1}^{C}p(X_c^N=x_c^N,Y^{C,N}=y^{C,N})\\
\propto& \prod_{c=1}^{C}\mathbb{P}(X_c^N=x_c^N,Y^{C,N}=y^{C,N}),
\end{aligned}$$

ce qui implique que

$$\underset{x^{C,N}\in\Omega_X^{C\times N}}{\arg\max}\ \mathbb{P}(X^{C,N}=x^{C,N},Y^{C,N}=y^{C,N}) = \underset{x^{C,N}\in\Omega_X^{C\times N}}{\arg\max}\prod_{c=1}^{C}\mathbb{P}(X_c^N=x_c^N,Y^{C,N}=y^{C,N})$$

On peut ensuite décomposer les maxima par chaîne.

$$\max_{x^{C,N}\in\Omega_X^{C\times N}}\prod_{c=1}^{C}\mathbb{P}(X_c^N=x_c^N,Y^{C,N}=y^{C,N}) = \prod_{c=1}^{C}\max_{x_c^N\in\Omega_X^N}\mathbb{P}(X_c^N=x_c^N,Y^{C,N}=y^{C,N})$$

Par conséquent on peut appliquer l'algorithme de Viterbi chaîne par chaîne. On pose

$$\delta_{c,n}(x_{c,n}) = \max_{x_c^{n-1}\in\Omega_X^{n-1}}\mathbb{P}(X_c^n=x_c^n,Y^{C,n}=y^{C,n}).$$

La fonction $\delta_{c,n}(x_{c,n})$ représente la suite d'états la plus probable dans la chaîne c jusqu'à l'état $x_{c,n}$. Il est possible d'écrire $\delta_{c,n}$ de façon récursive.

$$
\begin{aligned}
\delta_{c,n}(x_{c,n}) &= \max_{x_c^{n-1}\in\Omega_X^{n-1}} \mathbb{P}(X_c^n = x_c^n, Y^{C,n} = y^{C,n}) \\
&= \max_{x_c^{n-1}\in\Omega_X^{n-1}} \big[\mathbb{P}(X_c^{n-1} = x_c^{n-1}, Y^{C,n-1} = y^{C,n-1}) \\
&\quad \times \mathbb{P}(X_{c,n} = x_{c,n}, Y_n^C = y_n^C | X_c^{n-1} = x_c^{n-1}, Y^{C,n-1} = y^{C,n-1})\big] \\
&= \max_{x_{c,n-1}\in\Omega_X} \delta_{c,n-1}(x_{c,n-1})\mathbb{P}(X_{c,n} = x_{c,n}, Y_n^C = y_n^C | X_{c,n-1} = x_{c,n-1}, Y^{C,n-1} = y^{C,n-1}) \\
&= \max_{x_{c,n-1}\in\Omega_X} \delta_{c,n-1}(x_{c,n-1})\mathbb{P}(Y_n^C = y_n^C | X_{c,n-1} = x_{c,n-1}, Y^{C,n-1} = y^{C,n-1})A(x_{c,n-1}, x_{c,n}, y_n^C) \\
&= \max_{x_{c,n-1}\in\Omega_X} \big[\delta_{c,n-1}(x_{c,n-1})\phi(x_{c,n-1}, y_{c,n}, y_{n-1}^C) \\
&\quad \times \mathbb{P}(Y_n^{\mathcal{C}\setminus c} = y_n^{\mathcal{C}\setminus c} | X_{c,n-1} = x_{c,n-1}, Y^{C,n-1} = y^{C,n-1})A(x_{c,n-1}, x_{c,n}, y_n^C)\big] \\
&= \max_{x_{c,n-1}\in\Omega_X} \delta_{c,n-1}(x_{c,n-1})\phi(x_{c,n-1}, y_{c,n}, y_{n-1}^C)A(x_{c,n-1}, x_{c,n}, y_n^C)K_{c,n}
\end{aligned}
$$

On sait que $K_{c,n}$ est une constante. Il nous suffit donc de calculer $\tilde{\delta}_{c,n}(x_{c,n})$ définie par

$$\tilde{\delta}_{c,n}(x_{c,n}) = \max_{x_{c,n-1}\in\Omega_X} \tilde{\delta}_{c,n-1}(x_{c,n-1})\phi(x_{c,n-1}, y_{c,n}, y_{n-1}^C)A(x_{c,n-1}, x_{c,n}, y_n^C)$$

avec $\tilde{\delta}_{c,0} = \delta_{c,0}$. Afin d'obtenir la combinaison d'état la plus probable, il suffit de récupérer l'état $x_{c,n-1}$ en calculant de façon récursive $\gamma_{c,n}(x_{c,n})$:

$$\gamma_{c,n}(x_{c,n}) = \underset{x_{c,n-1}\in\Omega_X}{arg\max}\ \tilde{\delta}_{c,n-1}(x_{c,n-1})\phi(x_{c,n-1}, y_{c,n}, y_{n-1}^C)A(x_{c,n-1}, x_{c,n}, y_n^C)$$

5.5 Mise en oeuvre de l'algorithme Viterbi

— Initialisation : pour tout $x_{c,0} \in \Omega_X$, $\tilde{\delta}_{c,0}(x_{c,0}) = \varphi(y_{c,0})\pi(y_{c,0}, x_{c,0})$ et $\gamma_{c,0}(x_{c,0}) = 0$.
— Récursion pour tout $n \in \{0, \ldots, N\}$ et tout $c \in \{0, \ldots, C\}$.

$$
\begin{aligned}
\tilde{\delta}_{c,n}(x_{c,n}) &= \max_{x_{c,n-1}\in\Omega_X} \tilde{\delta}_{c,n-1}(x_{c,n-1})\phi(x_{c,n-1}, y_{c,n}, y_{n-1}^C)A(x_{c,n-1}, x_{c,n}, y_n^C) \\
\gamma_{c,n}(x_{c,n}) &= \underset{x_{c,n-1}\in\Omega_X}{\arg\max}\ \tilde{\delta}_{c,n-1}(x_{c,n-1})\phi(x_{c,n-1}, y_{c,n}, y_{n-1}^C)A(x_{c,n-1}, x_{c,n}, y_n^C)
\end{aligned}
$$

Une fois la récursion établie, il faut restaurer les état cachés. L'initialisation de la restauration est réalisée à partir de la fin de la chaîne. On peut restaurer l'état $X_{c,N}$ en regardant l'état qui maximise $\tilde{\delta}_{c,N}$. La restauration des états cachés antérieurs est faite de façon récursive.
— Initialisation pour tout $c \in \{0, \ldots, C\}$,

$$\hat{x}_{c,N} = \underset{x_{c,N}\in\Omega_X}{\arg\max} \tilde{\delta}_{c,n}(x_{c,n})$$

— Récursion pour tout $n \in \{0, \ldots, N\}$ et tout $c \in \{0, \ldots, C\}$..

$$\hat{x}_{c,n} = \gamma_{c,n}(\hat{x}_{c,n+1})$$

où $\hat{x}_{c,n}$ est l'état le plus probable de la variable $X_{c,n}$.

L'algorithme Viterbi a une complexité algorithmique de $O(CN|\Omega_X|)$.

5.6 Prédiction

Dans cette section, nous décrivons comment prédire l'état des variables observées de toutes les chaînes au temps $N+1$ à partir des observations. La prédiction au temps $N+1$ peut être faite chaîne par chaîne. On note $Pred(y_{c,N+1}) = \mathbb{P}(Y_{c,N+1} = y_{c,N+1} | Y^{C,N} = y^{C,N})$. Alors

$$\begin{aligned} Pred(y_{c,N+1}) &= \sum_{x_{c,N} \in \Omega_X} \mathbb{P}(Y_{c,N+1} = y_{c,N+1}, X_{c,N} = x_{c,N} | Y^{C,N} = y^{C,N}) \\ &= \sum_{x_{c,N} \in \Omega_X} \mathbb{P}(X_{c,N} = x_{c,N} | Y^{C,N} = y^{C,N}) \mathbb{P}(Y_{c,N+1} = y_{c,N+1} | X_{c,N} = x_{c,N}, Y^{C,N} = y^{C,N}) \\ &= \sum_{x_{c,N} \in \Omega_X} \rho_{c,N}(x_{c,N}) \phi(x_{c,N}, y_{c,N+1}, y_{c,N}) \end{aligned}$$

Il suffit ensuite de prendre l'argument maximal de $Pred(y_{c,N+1})$ pour trouver la prédiction de la flore levée au temps $N+1$ du champ c.

$$\hat{y}_{c,N+1} = \underset{y_{c,N+1} \in \Omega_Y}{\arg\max}\ Pred(y_{c,N+1})$$

La complexité algorithmique de l'algorithme de prédiction est en $O(C|\Omega_X|)$.

5.7 Discussion

Nous avons étendu les équations du Forward-Backward et détaillé les calculs de l'algorithme EM pour le MHMM-DF complet. Cela nous a permis de réduire la complexité algorithmique du EM d'exponentielle à linéaire en le nombre de chaînes. Le Forward-Backward légèrement modifié est un cas particulier de la propagation des croyances (en anglais : belief propagation Yedidia et al. (2003)) qui est souvent utilisée pour les réseaux bayésiens. La propagation des croyances permet de calculer la distribution marginale de chaque variable cachée conditionnellement aux variables observées. Comme pour le Forward-Backward, la propagation des croyances sur MHMM-DF peut être effectuée chaîne par chaîne car chaque chaîne cachée est indépendante d'une autre chaîne cachée sachant les données. Cette condition d'indépendance est vérifiée dans le cadre du MHMM-DF car les interactions entre les dynamiques locales proviennent uniquement de variables observées. Dans le cas où les interactions entre les dynamiques locales proviendraient aussi de variables cachées, le Forward-Backward ne pourrait être décomposé chaîne par chaîne et aurait une complexité exponentielle en le nombre de chaînes. Dans ce cadre, l'algorithme EM devrait être substitué pour une méthode moins coûteuse en complexité algorithmique comme le Variational Expectation Maximisation VEM Beal. (2003).

Chapitre 6

Expériences numériques

Dans cette section nous allons exposer les résultats obtenus à partir du sous-modèle du MHMM-DF associé à la dynamique des plantes annuelles sur données simulées.

Trois types d'expériences ont été réalisés à l'aide de données simulées. La première expérience consiste à évaluer la qualité des estimateurs. La deuxième est une sélection de modèle. Cela permet de déterminer le type de stratégie le plus probable pour un jeu de données parmi deux différents types de dynamiques. La dernière expérience évalue la qualité des prédictions et de la restauration des états des populations cachées. Ici, toutes les lois du MHMM-DF sont des Binomiales logistiques, définies aux paragraphes 4.4.2 p62 et 4.5.2.1 p65. Les simulations suivantes ont été effectuées avec 10 champs sur 100 pas de temps avec $|\Omega_X| = |\Omega_Y| = 5$. Seulement 10 simulation ont été effectuées par jeux de donnée en raison du temps de calcul important. En fixant la valeur des paramètres (τ, μ, ν), on simule les jeux de données des populations observables et non-observables. A partir des simulations, on estime la valeur des paramètres et on prédit et restaure l'état des populations observables au temps 101 et l'état de toutes les populations non-observables jusqu'au temps 100. Les méthodes ont été implémentées sous R et l'étape M utilise l'algorithme de limited-memory Broyden-Fletcher-Goldfarb-Shanno dans la fonction optim. Puisque l'algorithme EM trouve un maximum local, on lance 8 EM avec 8 jeux de paramètres initiaux générés aléatoirement et on garde les estimateurs avec la meilleure vraisemblance. L'algorithme EM s'arrête une fois que le nombre d'itérations arrive à 100 ou quand le maximum de la différence des estimateurs entre deux itérations est inférieur à 0.01.

6.1 Qualité d'estimation

La qualité des estimateurs est évaluée selon leur variance et leur biais. 14 jeux de paramètres ont été utilisés et un seul des paramètres varie de 0 à 6.5. Les autres paramètres sont fixés et leur valeur est indiquée dans chaque cas dans le titre des figures. On a choisi de représenter les figures 6.1, 6.2 et 6.3 quand ν_1 varie car les figures sont similaires quand on fait varier ν_2 et ν_3. On a cependant aussi représenté en figure 6.4 la variance des estimateurs quand ν_3 varie en fixant $\nu_1 = 4$, valeur pour laquelle la variance pour tous les estimateurs est faible en figure 6.1.

Pour chaque jeu de paramètres, on simule 10 trajectoires et un jeu d'estimateurs est calculé pour chacune des trajectoires. On calcule ensuite la variance et le biais de chaque estimateur à partir de ces 10 estimations. On remarque dans la figure 6.1 que la variance des paramètres est grande quand le paramètre ν_1, associé à la survie des populations cachées, est grand ou petit. Notons que pour toute valeur de ν_1, les paramètres ν_0 et μ_0 sont bien estimés.

La figure 6.3 représente la fréquence des états des populations cachées quand ν_1 varie. Quand ν_1 est petit, l'état le plus fréquenté est l'état d'extinction de la population cachée. Inversement, quand ν_1 est grand, l'état le plus fréquenté est l'état maximal de la population cachée. Quand

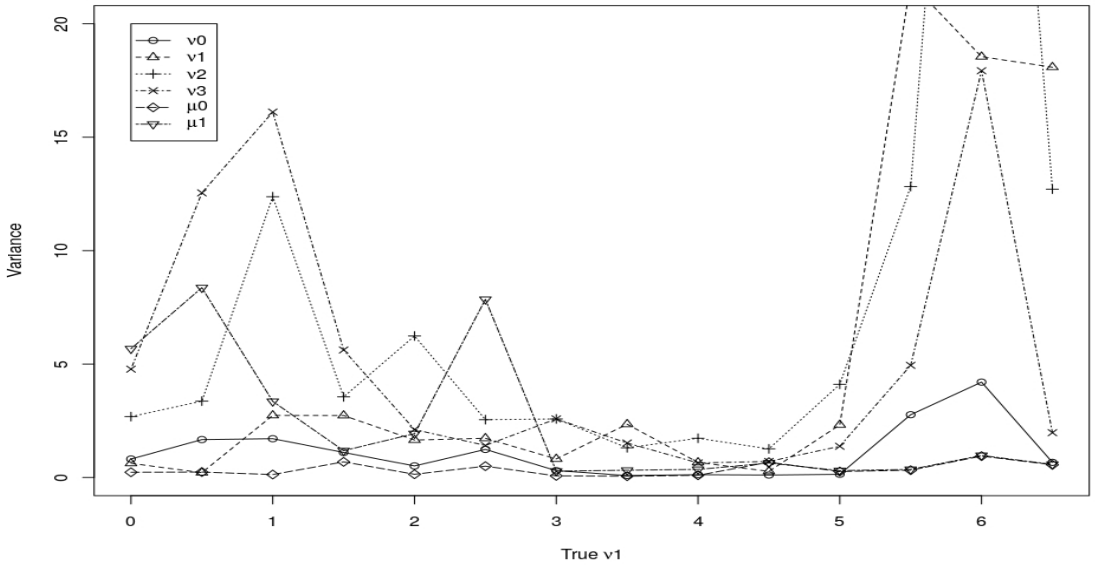


FIGURE 6.1 – Évolution de la variance des estimateurs avec $(\tau, \mu_0, \mu_1, \nu_0, \nu_2, \nu_3) = (-1, -3.7, 6.5, -3, 4, 2)$ où ν_1 varie de 0 a 6.5 avec un pas de 0.5.

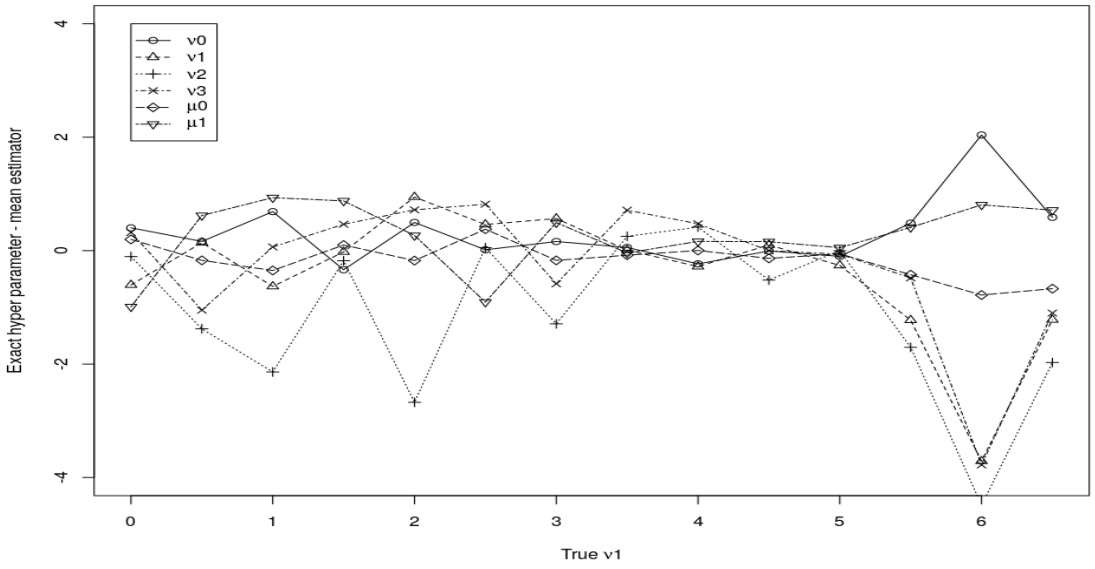


FIGURE 6.2 – Évolution de la différence entre les paramètres exacts et les estimateurs quand les paramètres exacts sont $(\tau, \mu_0, \mu_1, \nu_0, \nu_2, \nu_3) = (-1, -3.7, 6.5, -3, 4, 2)$ où ν_1 varie de 0 a 6.5 avec un pas de 0.5.

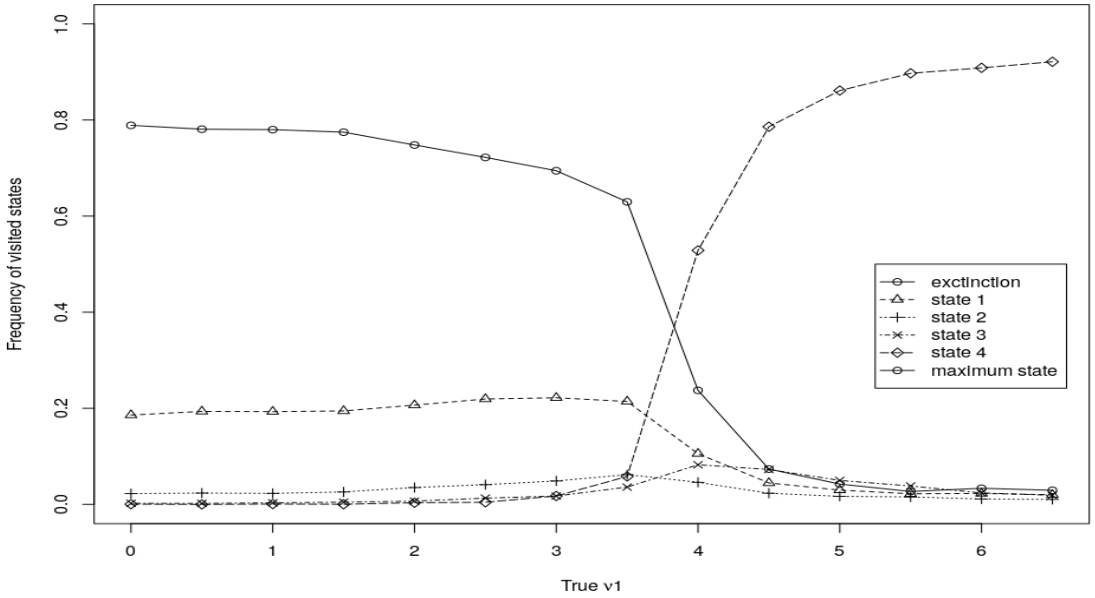


FIGURE 6.3 – Évolution de la fréquence de l'état des populations cachées avec $(\tau, \mu_0, \mu_1, \nu_0, \nu_2, \nu_3) = (-1, -3.7, 6.5, -3, 4, 2)$ où ν_1 varie de 0 jusqu'à 6.5 avec un pas de 0.5.

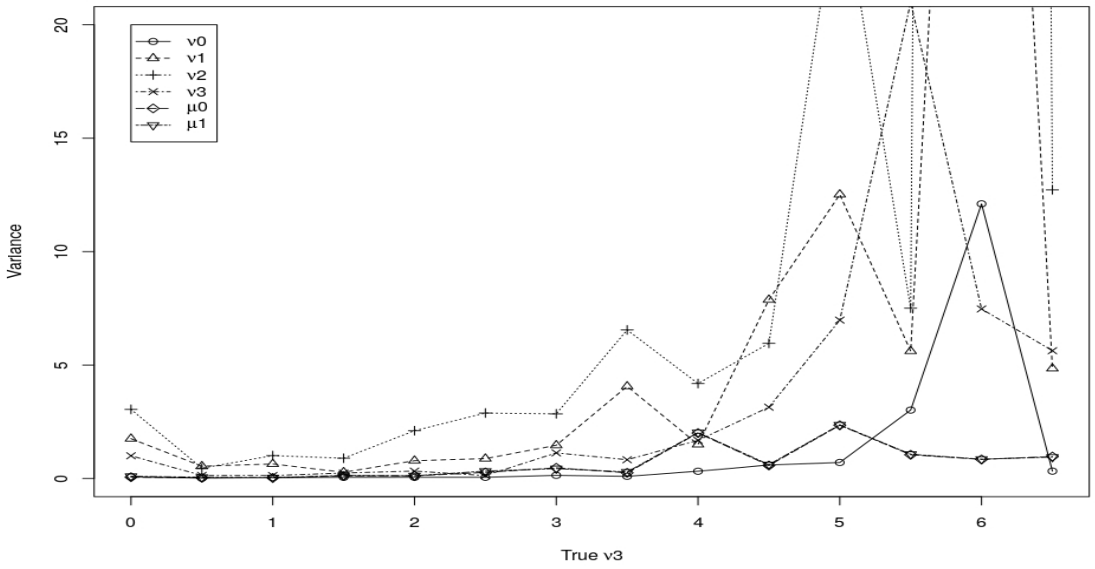


FIGURE 6.4 – Évolution de la variance des estimateurs avec $(\tau, \mu_0, \mu_1, \nu_0, \nu_1, \nu_2) = (-1, -3.7, 6.5, -3, 4, 4)$ où ν_3 varie de 0 a 6.5 avec un pas de 0.5.

$\nu_1 \in [3, 5]$, les estimateurs ont une variance faible. Autrement dit, quand il n'y a pas d'état qui prédomine largement sur les autres, la dynamique de l'espèce peut facilement être estimée.

On remarque que la variance de μ_1 est grande quand ν_1 est petit, c'est-à-dire quand l'état d'extinction de la population cachée prédomine largement. Ce résultat n'est pas étonnant car la contribution de μ_1 dans la probabilité d'émission est nulle quand la population cachée est éteinte. Par conséquent, moins de données sont utilisées dans l'estimation de μ_1, induisant une forte variance.

La figure 6.4 représente le graphe des variances des estimateurs avec différents jeux de paramètres. Dans cette figure, le paramètre de la survie des populations cachées est fixé à 4 et le paramètre de colonisation varie entre 0 et 6.5. On remarque que la variance des estimateurs ν_1 et ν_2 est grande quand ν_3 est grand. La figure 6.2, qui correspond au biais des estimateurs avec ν_1 variable, montre que quand $\nu_1 \in [3, 5]$ les estimateurs ont un biais plus petit que pour les petites et les grandes valeurs du paramètre ν_1.

Quand les paramètres de la probabilité de transition de la population cachée sont trop grands, la population cachée ne fréquente que l'état maximal. Inversement, quand les paramètres de la probabilité de transition de la population cachée sont trop faibles, les processus n'influencent pas assez la population cachée et l'état le plus fréquenté est alors l'état d'extinction. Dans ces deux cas extrêmes, déterminer la contribution de chaque processus impliqué est difficile car différents vecteurs de paramètres sont susceptibles de générer les mêmes données.

6.2 Sélection de modèle

La sélection de modèle permet d'établir le modèle qui représente le mieux les données et ainsi d'éventuellement déterminer si un processus donné est négligeable dans la dynamique de l'espèce.

On veut comparer dans un premier temps les dynamiques avec ou sans survie de la population cachée puis les dynamiques avec ou sans colonisation. Le modèle paramétrique du sous-modèle du MHMM-DF associé à la dynamique des plantes annuelles, que l'on appellera le modèle adventice complet, a 7 paramètres tandis que les modèles sans colonisation ou sans survie de la banque de graines n'en ont que 6. Pour les comparer, on simule des trajectoires à l'aide du modèle adventice complet en faisant varier le paramètre associé à la survie de la population cachée ou à la colonisation. Ces trajectoires fournissent des données sur lesquelles on estime les paramètres pour les deux modèles. Pour choisir le plus adéquat des deux, on a décidé d'utiliser le critère d'information de Akaike (Akaike (1981)Akaike (2011)), noté AIC en abrégé. L'AIC est calculé pour chaque modèle et le modèle avec l'AIC le plus petit est sélectionné. Néanmoins, on estime qu'une différence d'AIC inférieure à 1 est insuffisante pour départager deux modèles.

Afin de choisir entre les modèles avec ou sans survie de la banque de graines, on a simulé 30 trajectoires avec 100 pas de temps sur 10 chaînes avec les paramètres suivants : $\mu = (-3.7, 6.5)$, $\tau = -1$ et $\nu = (-3, \nu_1, 4, 2)$, où ν_1, le paramètre associé à la survie de la banque de graines, varie dans $\{0, 1, 2, 3\}$. Fixer ν_1 égal à 0 est équivalent à supposer la survie inexistante. On s'attend donc à ce que le modèle adventice complet, donc avec survie, soit choisi à celui sans survie quand ν_1 grandit. Pour ν_1 égal à 0, 1, 2 et 3, le modèle sans survie de la banque de graines a été choisi respectivement 24, 25, 17 et 4 fois sur 30. Ces résultats ne sont pas satisfaisants puisque le modèle sans survie a été sélectionné plus de fois que le modèle adventice complet pour ν_1 égal 1 ou 2. Ce phénomène peut être dû à la forte variance des estimateurs quand $\nu_1 = 0, 1, 2$ (voir figure 6.1). Pour valider cette hypothèse, on refait les simulations précédentes, avec les mêmes paramètres, à l'exception de ν_2, fixé à 8 contre 4 précédemment. On choisit ν_2 plus grand puisque des simulations, non données ici, ont montré que la variance des estimateurs est plus petite pour $\nu_2 = 8$ que pour $\nu_2 = 4$. Pour $\nu_2 = 8$ et ν_1 égal à 0, 1, 2 et 3, le modèle sans survie de la banque de graines a été choisi respectivement 22, 8, 1 et 1 fois sur 30. Ces résultats sont plus satisfaisants et confirment l'hypothèse selon laquelle une mauvaise variance des estimateurs des paramètres induit une mauvaise sélection de modèle.

Afin de sélectionner un modèle avec ou sans colonisation, nous avons simulé 30 trajectoires avec les paramètres suivants : $\mu = (-3.7, 6.5)$, $\tau = -1$ et $\nu = (-3, 4, 4, \nu_3)$, où ν_3, le paramètre associé à la colonisation, varie dans $\{0, 1, 2, 3\}$. Fixer ν_3 égal à 0 revient à considérer un modèle sans colonisation. Avec ν_3 égal à 0, 1, 2 et 3, le modèle sans colonisation a été choisi respectivement 29, 19, 4 et 0 fois sur 30. Comme attendu, quand ν_3 grandit, c'est-à-dire quand l'influence de la colonisation augmente, le modèle adventice complet est préféré à celui sans colonisation.

6.3 Restoration et prédiction

Nous nous intéressons ici à la moyenne des pourcentages des états des populations cachées correctement restaurés et à la moyenne des pourcentages des états des populations observées correctement prédites. On se donne un vecteur de paramètres et on simule 10 trajectoires sur 10 patchs sur 100 années avec $|\Omega_X| = |\Omega_Y| = 5$. De plus, on simule 100 fois les populations observables sur les 10 patchs au temps 101. Ensuite, on utilise les 100 premières variables observées sur les 10 patchs pour estimer la valeur des paramètres du modèle adventice complet. Une fois les estimateurs des paramètres obtenus, on les utilise pour la prédiction des populations observables du 101^e pas de temps et la restauration des populations cachées des 100 premiers pas de temps. On compare la prédiction obtenue avec les 100 simulations du 101^e pas de temps. 28 différents vecteurs de paramètres ont été utilisés. Les 14 premiers vecteurs de paramètres sont de la forme $\mu = (-3.7, 6.5)$, $\tau = -1$ et $\nu = (-3, \nu_1, 4, 2)$ avec le paramètre associé à la survie des graines dans la banque de graines ν_1 qui varie dans $\{0, 0.5, 1, \ldots, 6.5\}$. Les 14 autres vecteurs de paramètres sont de la forme $\mu = (-3.7, 6.5)$, $\tau = -1$ et $\nu = (\nu_0, 4, 4, 2)$ avec ν_0 qui varie dans $\{-6.5, -6, -5.5, \ldots, 0\}$.

Les figures 6.5 et 6.6 montrent que la qualité de l'estimation des variables cachées est souvent supérieure à 70% et est d'autant meilleure que ν_1 augmente. En revanche, la qualité de la prédiction est basse quand ν_1 est grand. La figure 6.6 montre que quand la colonisation extérieure diminue,

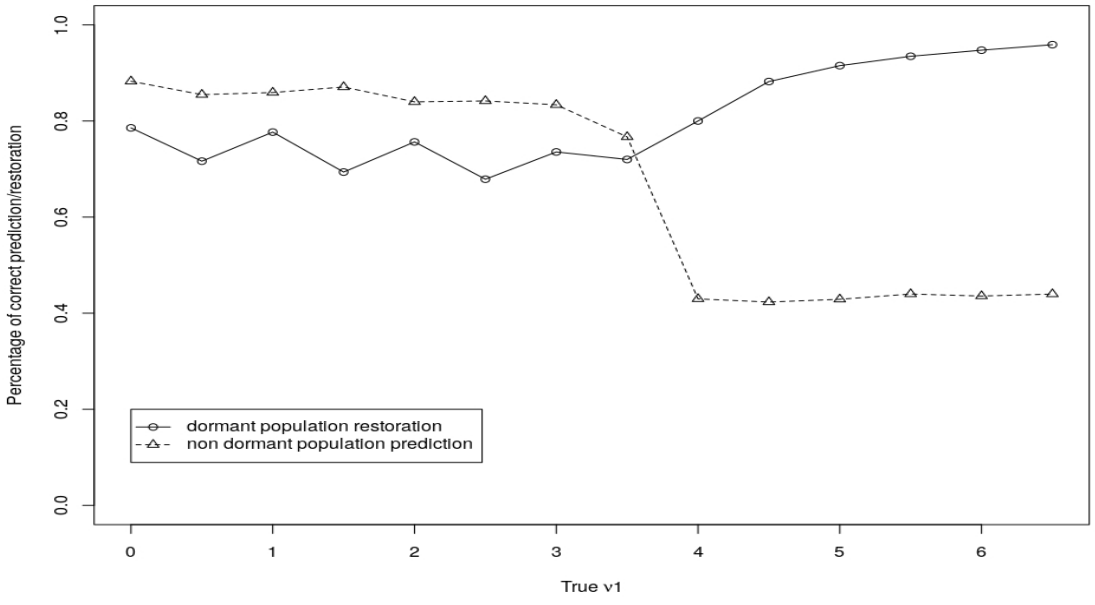


FIGURE 6.5 – Moyenne des pourcentages des populations correctement restaurées et prédites pour ν_1 variant de 0 à 6.5.

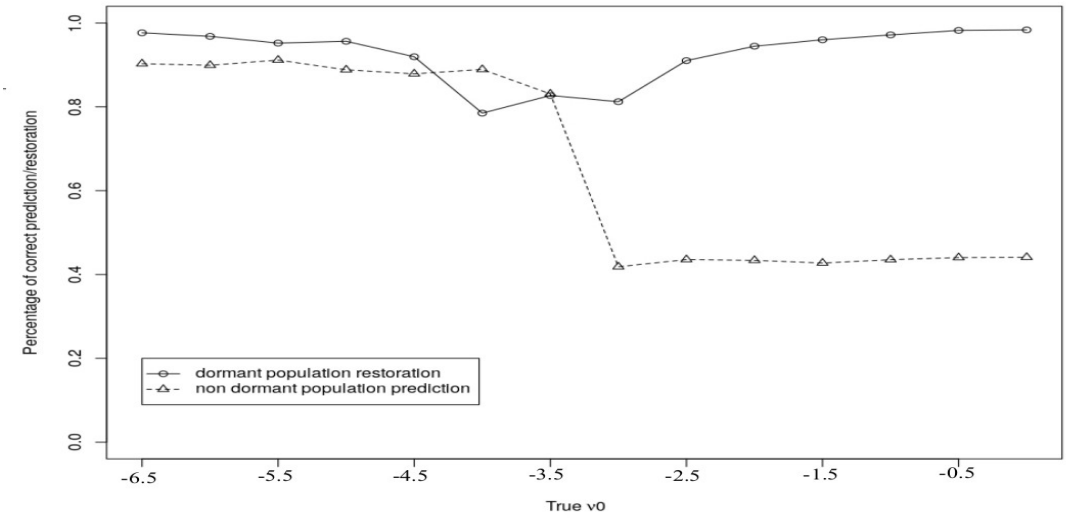


FIGURE 6.6 – Moyenne des pourcentages des populations correctement restaurées et prédites pour ν_0 variant de -6.5 à 0.

c'est-à-dire quand ν_0 diminue, la restauration s'améliore. Par conséquent, il semblerait que la colonisation extérieure crée un bruit dans la prédiction et la restauration pour de faibles valeurs de ν_1. Pour de grandes valeurs de ν_1, les populations cachées visitent fréquemment l'état maximum et le pourcentage de restauration correcte augmente. Même si distinguer la contribution de chaque processus est de plus en plus difficile quand ν_1 grandit, la variance de la loi Binomiale associée à l'état de la banque de graines décroît.

En revanche, le pourcentage de prédiction correcte diminue jusqu'aux alentours de 40% quand $\nu_1 \geq 4$. On montre facilement que quand ν_1 grandit, la variance de la loi Binomiale associée à la germination augmente. De plus, comme l'illustre la figure 6.3, la population cachée fréquente majoritairement l'état maximal. Quand la population cachée est dans l'état maximal, la population observée a une probabilité de 44% de prendre l'état maximal, 39% de prendre l'état 3, 13% de prendre l'état 2, 2% de prendre l'état 1 et 0.1% de ne pas germer. Puisque le modèle prédit l'état le plus probable de la population observée, il n'est pas étonnant que 40% seulement des prédictions soient justes.

On peut donc en déduire que la qualité de la restauration et de la prédiction est d'autant meilleure que la variance de la loi Binomiale associée à la population considérée diminue.

Afin d'illustrer les capacités de l'algorithme EM, la même expérience a été reproduite avec 10 pas de temps et 100 champs pour 30 simulations. Pour $\mu = (-3.7, 6.5)$, $\tau = -1$ et $\nu = (-3, 3, 4, 2)$, on a obtenu 80% de prédictions correctes des populations observables et 68.9% de restaurations correctes des populations cachées.

6.4 Expériences numériques avec dépendance à la culture

Dans le chapitre suivant, on souhaite utiliser notre modèle pour analyser les données d'Epoisses. Jusque-là, toutes les lois du modèle adventice complet étaient de type Binomiale logistique, loi définie aux paragraphes 4.5.2.1 p65 et 4.4.2 p62. Modéliser la loi d'émission par une loi Binomiale

logistique autorise une colonisation extérieure vers des populations observées, ce qui est impossible dans le cadre de la dynamique des plantes. Pour pallier ce problème, on choisit de modéliser la loi d'émission par une loi Binomiale uniquement pour les états cachés non-éteints, définie au paragraphe 4.4.4 p63.

Dans le cadre des données d'Epoisses, il y a une possible rotation de cultures. Sur un même patch, la culture peut changer d'une année sur l'autre. Cependant, chaque culture a un jeu de paramètres qui lui est propre. On estime ici les paramètres du MHMM-DF pendant $N = 14$ années sur $C = 90$ patchs présentant deux cultures différentes.

Simulation	Paramètres exacts	Estimation
$\begin{pmatrix}\nu_0\\ \nu_1\\ \nu_2\\ \nu_3\end{pmatrix}_{cult1}$	$\begin{pmatrix}-3\\ 4\\ 3\\ 2\end{pmatrix}$	$\begin{pmatrix}-2.895\\ 4.818\\ 1.788\\ 1.747\end{pmatrix}$
$\begin{pmatrix}\nu_0\\ \nu_1\\ \nu_2\\ \nu_3\end{pmatrix}_{cult2}$	$\begin{pmatrix}-3\\ 0\\ 5\\ 3\end{pmatrix}$	$\begin{pmatrix}-2.826\\ 0\\ 5.187\\ 4.310\end{pmatrix}$
$\begin{pmatrix}\mu_0\\ \mu_1\end{pmatrix}_{cult1}$	$\begin{pmatrix}-3.7\\ 7\end{pmatrix}$	$\begin{pmatrix}-3.693\\ 6.476\end{pmatrix}$
$\begin{pmatrix}\mu_0\\ \mu_1\end{pmatrix}_{cult2}$	$\begin{pmatrix}-3.7\\ 5\end{pmatrix}$	$\begin{pmatrix}-4.376\\ 5.660\end{pmatrix}$
τ	-1	-0.631
Restauration BG		75.6%

Le tableau montre qu'en dépit de cette contrainte supplémentaire, les estimations restent bonnes.

6.5 Discussion

Nous avons établi qu'il est facile de déterminer le modèle le plus adéquat aux données ainsi que la contribution de chaque processus impliqué, hormis dans des cas où les variables cachées ne fréquentent que des états extrêmes. Les résultats laissent présager que la qualité de la prédiction et de la restauration s'améliore avec la diminution de la variance de la loi Binomiale associée à la population considérée.

Chapitre 7

Analyse des données d'Epoisses

Nous allons ici estimer la dynamique des adventices du complexe d'Epoisses grâce aux données transmises par Stéphane Cordeau de l'INRA de Dijon. Ces estimations nous permettent de mieux comprendre la part relative de chaque processus dans la dynamique de l'adventice. L'analyse des données est réalisée grâce à trois modèles, un MHMM-DF avec dépendance à la saison de la culture, un MHMM-DF sans dépendance à la saison et le modèle de Pluntz et al. (2018). Les résultats selon ces différents modèles sont analysés puis comparés en étudiant successivement les estimations de la colonisation, de la survie des graines dans le sol, de la germination et de la probabilité de non-dispersion des graines venant des plantes adultes. Enfin, on s'intéresse à la qualité de la prédiction de la population observable et de la restauration de la population cachée.

Toutes les estimations dans cette section supposent 5 états possibles pour la flore levée et 5 états possibles pour la banque de graines. Les états de la flore levée sont définis dans la section 7.2. La fonction d'agrégation moyennée a été utilisée afin d'agréger les données des champs voisins. Les probabilités du modèle sont déterminées par (π, A, ϕ), où π est la probabilité initiale, A est la matrice de transition et ϕ est la matrice d'émission.

$$\begin{aligned} A(x_{c,n}, x_{c,n+1}, y^C_{n+1}) &= \mathbb{P}(X_{c,n+1} = x_{c,n+1} | X_{c,n} = x_{c,n}, Y^C_{n+1} = y^C_{n+1}) \\ \phi(x_{c,n}, y_{c,n+1}) &= \mathbb{P}(Y_{c,n+1} = y_{c,n+1} | X_{c,n} = x_{c,n}) \\ \pi(x_{c,0}) &= \mathbb{P}(X_{c,0} = x_{c,0}) \end{aligned}$$

De plus, la loi d'émission, qui régit la germination dans le MHMM-DF pour la dynamique des adventices, est une loi Binomiale logistique uniquement pour les états cachés non-éteints (voir 4.4.4). La loi de transition de la banque de graines et la loi initiale sont des lois Binomiales avec la fonction logistique (voir 4.5.2.1).

7.1 Les données d'Epoisses

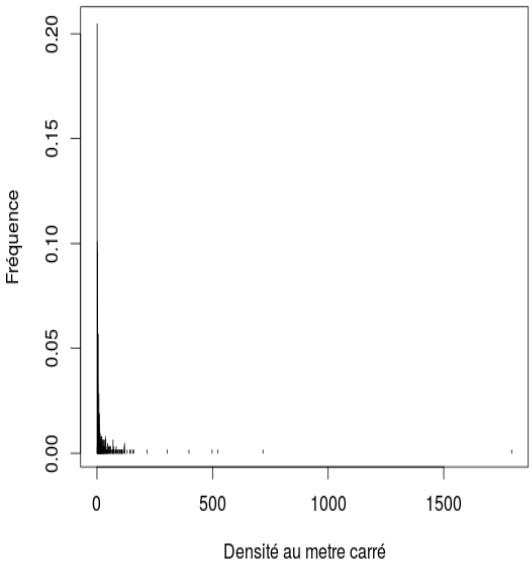
Les études menées sur le complexe d'Epoisses ont pour objectif de concevoir et d'expérimenter des systèmes de culture zéro pesticide dans différentes situations de production, et d'en évaluer les performances agronomiques, économiques, environnementales et sociales.

Les données d'Epoisses comprennent des données sur la flore levée des adventices de 2000 à 2017. Le complexe d'Epoisses est composé de deux groupes de cinq champs chacun. Chacun des dix champs est découpé en neuf zones, appelées patchs, sur lesquelles les données sont récoltées sur la flore levée. Les données sur la flore levée des adventices ont été récoltées au mètre carré soit un mois avant, soit pendant le semis de la culture, soit après le semis au moment de la levée de la culture. Cependant, les données de la flore levée sont manquantes pour deux zones de l'année 2003. Les données par zones sont faites à partir d'une moyenne sur quatre quadrats dont la position

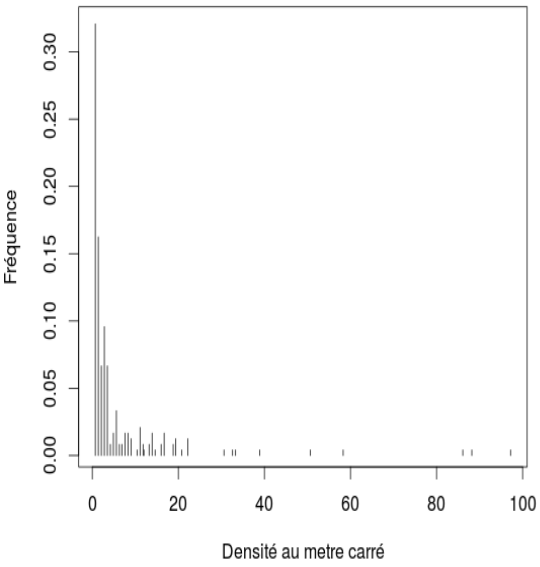
varie d'année en année. De plus, des données sur la banque de graines sont récoltées en interculture sur une zone par champ. Néanmoins, les données sur la banque de graines n'ont pas été collectées toutes les années. 70 espèces d'adventices ont été identifiées lors de la collecte de données.

Disposition des champs du complexe d'Epoisses

Dans cette section nous allons nous concentrer sur les 7 espèces d'adventices *Chenopodium album*, *Solanum nigrum*, *Alopecurus myosuroides*, Fallopia convolvulus, *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare*. Les analyses préliminaires à l'estimation sur les adventices ne sont proposées que pour les espèces *Chenopodium album*, *Solanum nigrum*, *Alopecurus myosuroides* et *Fallopia convolvulus* afin de ne pas surcharger le document. Par la suite, nous déterminons la probabilité de survie de la banque de graines, de germination des graines et la probabilité de colonisation pour les 7 espèces. Ces espèces ont été choisies pour plusieurs raisons. Tout d'abord, elles sont abondantes dans les champs agricoles d'Epoisses. A l'exception de *Aethusa cynapium*, ces espèces ont déjà été étudiées grâce au modèle de Pluntz et al. (2018) sur les données Biovigilance. Ainsi, il sera possible de comparer les résultats des deux études. Toutes les espèces considérées sont dormantes d'après Baskin and Baskin (1998). Les figures 7.1 présentent les histogrammes des densités au mètre carré des espèces *Chenopodium album*, *Solanum nigrum*, *Alopecurus myosuroides* et *Fallopia convolvulus*. Les densités nulles ont été écartées pour une meilleure lisibilité. Ces figures révèlent que les valeurs faibles de densité ont une plus forte fréquence d'observation. Le nombre de densités non nulles pour la flore levée des espèces *Alopecurus myosuroides*, *Chenopodium album*, *Fallopia convolvulus*, *Solanum nigrum*, *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare* sont respectivement 646, 252, 736, 275, 547, 406, 418 sur 1530 données.

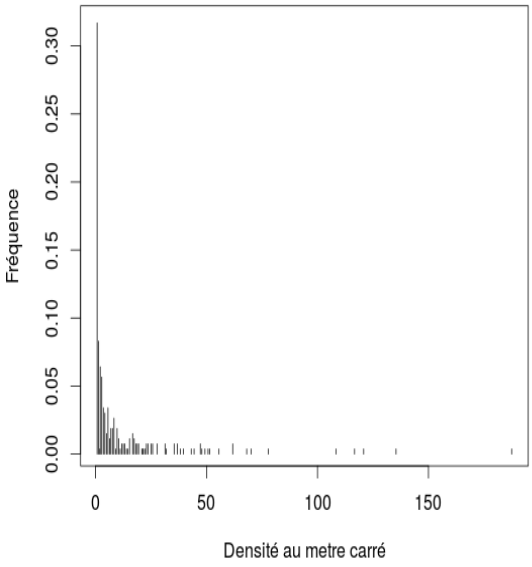


(a) *Alopecurus myosuroides*

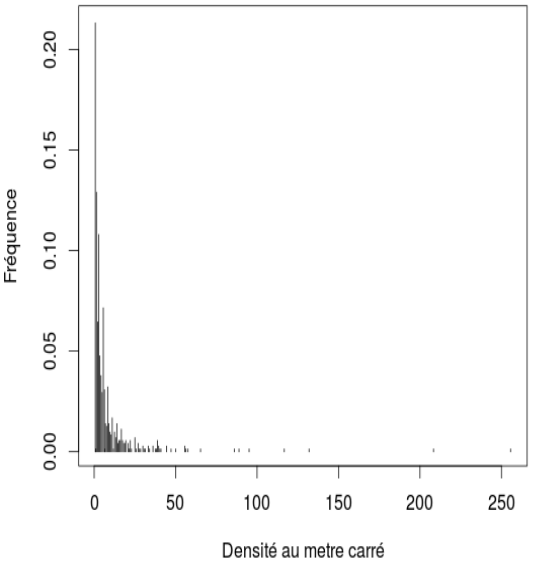


(b) *Chenopodium album*

FIGURE 7.1 – Histogramme des fréquences des densités au mètre carré non nulles des espèces



(a) *Solanum nigrum*



(b) *Fallopia convolvulus*

FIGURE 7.2 – Histogramme des fréquences des densités au mètre carré non nulles des espèces

L'une des particularités du complexe agricole d'Epoisses se trouve dans la diversification des méthodes de désherbage employées par champ. Dans chaque groupe de 5 champs, on distingue 5 stratégies de gestion, à savoir une agriculture raisonnée $S1$, une stratégie comportant un travail du sol $S2$, un désherbage mécanique $S3$, un désherbage mécanique et chimique $S4$ et un désherbage chimique $S5$. Chaque méthode de gestion est appliquée à deux champs qui ne sont pas dans le même groupe. Le tableau 7.1 rassemble le nombre d'observations non nulles par système de désherbage par espèce concernée.

	S1	S2	S3	S4	S5
Alopecurus myosuroides	77.00	183.00	105.00	139.00	142.00
Chenopodium album	4.00	33.00	56.00	83.00	76.00
Solanum nigrum	2.00	41.00	50.00	74.00	108.00
Fallopia convolvulus	120.00	82.00	153.00	173.00	208.00

TABLE 7.1 – Tableau du nombre d'observations par type de désherbage par type d'adventice dans le jeu de données d'Epoisses

On remarque que dans les champs qui appliquent l'agriculture raisonnée, les adventices sont moins présentes que dans les champs avec d'autres types de stratégies. On note aussi que les stratégies intégrant un désherbage chimique sont en moyenne moins efficaces que les autres.

Pour intégrer les méthodes de gestion des adventices au modèle, une première idée serait d'estimer les paramètres pour chacun des types de gestion. Néanmoins, cela réduirait grandement le nombre de données sur lesquelles reposerait l'estimation, ce qui nuirait à sa qualité. Une autre idée serait de choisir un compromis en regroupant les types de gestion en 2 familles, par exemple. Cependant, il faudrait déterminer les critères selon lesquels on les regrouperait. La stratégie de gestion $S2$, seule stratégie qui ne consiste pas à désherber, semble difficilement combinable avec toute autre stratégie de gestion et on possède peu de données sur des champs utilisant cette stratégie pour avoir une estimation précise. On a donc choisi de ne pas tenir compte de la méthode de gestion des adventices dans l'analyse des données d'Epoisses.

Pendant les 17 années de l'étude, 19 types de cultures différentes ont été mis en place. Il est important de souligner que les données sur une même adventice présente dans des cultures différentes ne sont pas récoltées au même moment dans l'année. De plus, d'après Donohue (2005) la germination des graines est fortement influencée par la saison. Ainsi la densité d'adventices risque de varier selon la culture. Au lieu de regarder les densités des adventices dans chaque type de culture, nous avons séparé les cultures en deux groupes : les cultures d'hiver et d'été. Les histogrammes 7.3, 7.4 illustrent les densités des espèces dans les cultures d'hiver et les cultures d'été.

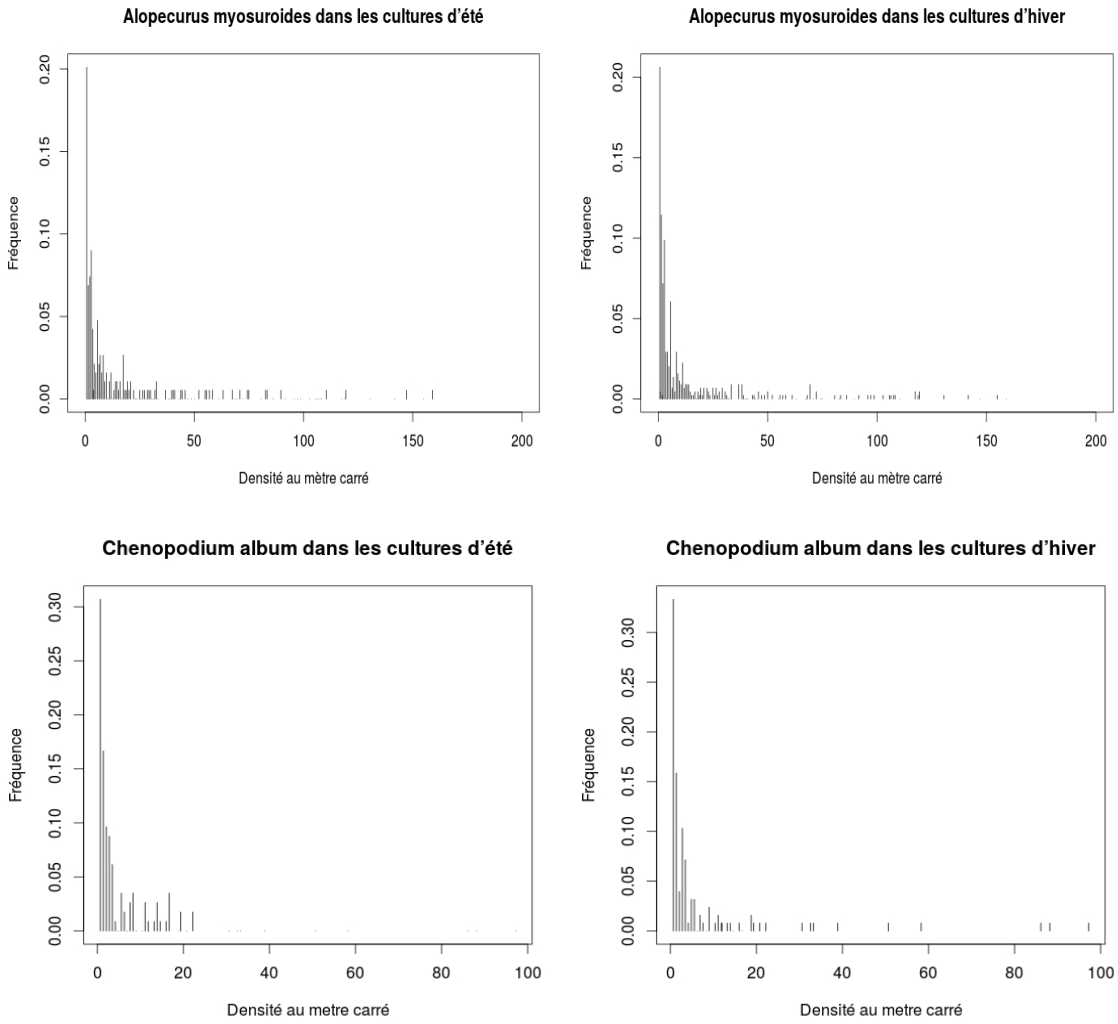


FIGURE 7.3 – Histogramme des observations non-nulles dans le jeu de données d'Epoisses pour les espèces *Alopecurus myosuroides*, *Chenopodium album* dans les cultures d'hiver et d'été

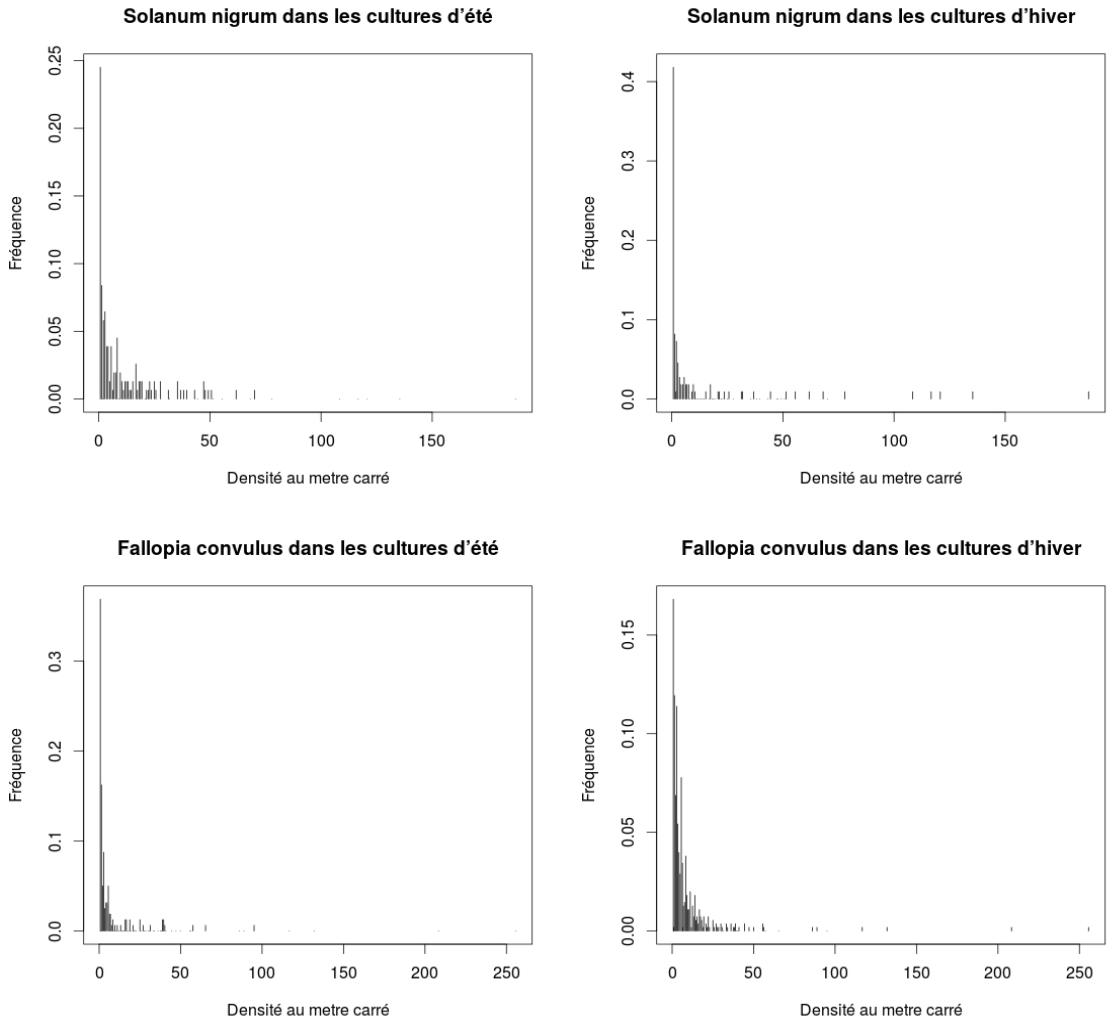


FIGURE 7.4 – Histogramme des observations non-nulles des espèces *Solanum nigrum* et *Fallopia convolvulus* dans les cultures d'hiver et d'été

Ces histogrammes mettent en évidence que *Alopecurus myosuroides* et *Fallopia convolvulus* sont plus présentes pendant l'hiver. Les moyennes de la densité des observations sur une culture d'hiver sont respectivement $9.06m^{-2}$, $0.86m^{-2}$, $1.36m^{-2}$ et $4.17m^{-2}$ pour les espèces *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum* et *Fallopia convolvulus*. Les moyennes de la densité des observations sur une culture d'été sont respectivement $7.66m^{-2}$, $1.23m^{-2}$, $3.93m^{-2}$ et $2.46m^{-2}$ pour les espèces *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum* et *Fallopia convolvulus*. La dynamique des adventices dépend de la saison. Ainsi, il serait judicieux d'utiliser une version du modèle MHMM-DF avec des paramètres variant selon la saison de la culture. Même si, comme pour la stratégie de gestion, la qualité de l'estimation subit une dégradation, on estime la dynamique des adventices selon la saison. Il y a 1109 données sur les cultures d'hiver et 421 données sur celles d'été. Au final, pour chacune de ces 4 espèces et par saison de la culture, au moins 114 données non-nulles ont été récoltées.

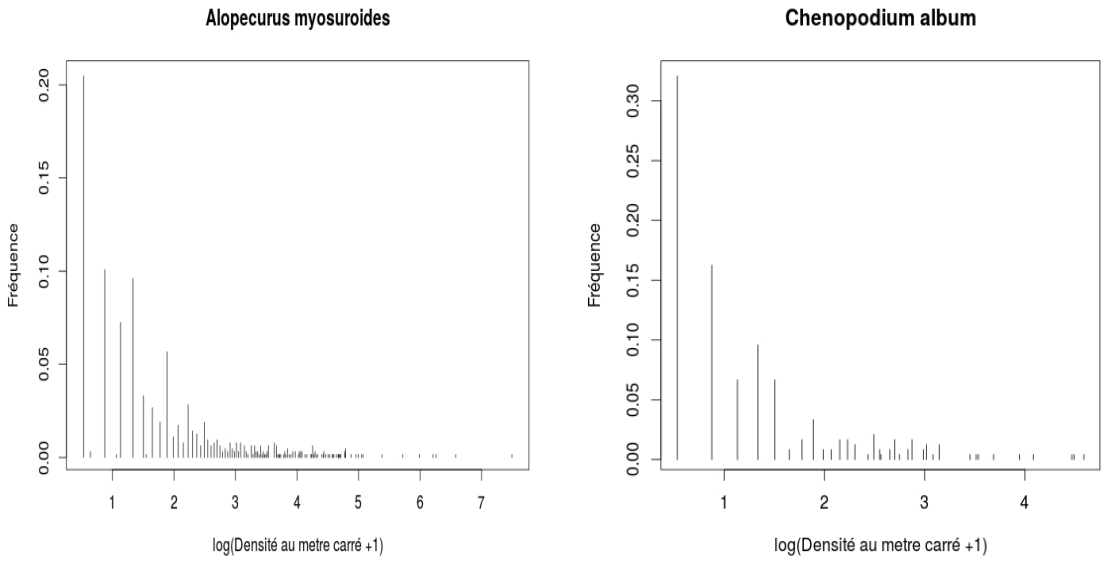
7.2 Comment définir les classes d'abondance ?

La section suivante est consacrée à la création des classes d'abondance à partir des données. Une façon de définir les classes d'abondance à partir des données de densité serait à l'aide de l'échelle Barralis. L'échelle Barralis est souvent utilisée dans l'échantillonnage de plantes. Le tableau ci-dessous représente le découpage en classes de la densité de la flore levée pour cette échelle.

plantes au mètre carré	classe associée
vue une fois sur l'aire d'observation	1
$d < 0.1$	2
$0.1 < d < 1$	3
$1 \leq d < 3$	4
$3 \leq d < 10$	5
$10 \leq d < 20$	6
$20 \leq d < 50$	7
$50 \leq d$	8

Plusieurs problèmes se posent en prenant une échelle comme celle-ci pour créer des classes d'abondance. Premièrement, les classes définies ci-dessus n'incluent pas l'état d'extinction de la population observable. Deuxièmement les classes d'abondance sont indépendantes des données sur les espèces. Ainsi, les effectifs ne sont pas forcément répartis uniformément à travers les classes. Cela peut conduire à une représentation grossière de la distribution de l'espèce.

Avant de présenter un autre découpage en classes, nous appliquons une transformation logarithmique. Elle permet de rapprocher des valeurs extrêmes pour obtenir des graphes de distribution moins étendus. Elle est particulièrement efficace pour normaliser les distributions désaxées vers la droite, comme dans le cas des données d'Epoisses dans les figures 7.1. En effet le coefficient d'asymétrie, aussi appelé moment d'ordre trois, des données sur l'espèce *Alopecurus myosuroides* est de 22.7, ce qui confirme une distribution désaxée vers la droite. Il est conseillé d'ajouter 1 à chaque observation avant d'effectuer la transformation logarithmique, pour éviter d'augmenter l'écart entre les observations comprises initialement entre 0 et 1.



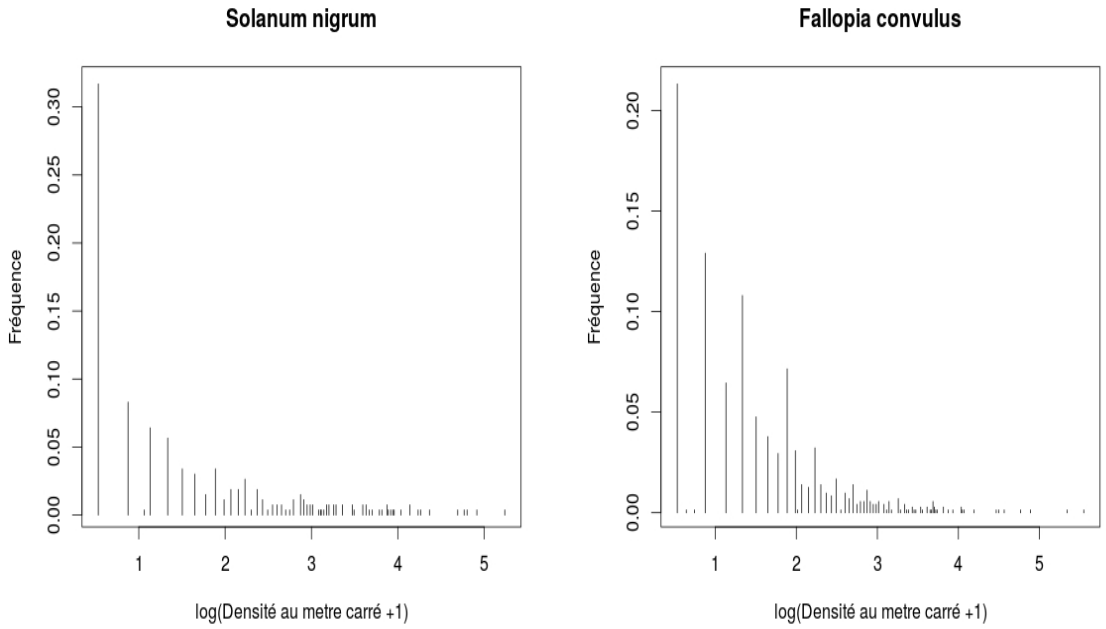


FIGURE 7.5 – Histogramme du log des observations non-nulles des espèces

La figure 7.5 illustre la transformation logarithmique des données. Les densités nulles ont été retirées pour mieux visualiser l'histogramme.

Regardons maintenant comment créer un découpage en classes à partir des données transformées. Pour éviter une représentation grossière de la distribution de l'espèce en classes, il existe plusieurs méthodes pour découper des données en classes d'abondance. Certaines méthodes cherchent à avoir des classes homogènes en minimisant la variance intra-classes tout en maximisant la variance inter-classes. D'autres méthodes s'intéressent à l'équirépartition des données à travers les classes ou bien au nombre minimal de données par classe lors du découpage en classes. Cependant, une répartition équitable des données ne correspond pas forcément à la meilleure représentation possible. Dans le cas des données d'Epoisses, les populations observées sont majoritairement dans l'état d'extinction. De ce fait, l'équirépartition des données en classes ne peut pas être respectée. Nous avons donc créé une classe spécifique pour les populations observées éteintes. Puis, nous avons recherché un découpage en classes des observations non nulles. L'une des méthodes que nous avons cherché à appliquer consiste à utiliser la divergence de Kullback-Leibler. On cherche une fonction de densité en escalier qui minimise la divergence de KullbackLeibler avec la fonction de densité associée aux données. Toutefois, avec cette méthode, les classes n'ont pas forcément la même taille d'intervalles et les intervalles ainsi choisis peuvent ne pas contenir d'observation, ce qui peut, en théorie, nuire à l'identifiabilité du modèle et à la variance des estimateurs. Ainsi, il est préférable d'avoir des observations dans toutes les classes d'abondance.

Pour déterminer les classes d'abondance, nous avons choisi une méthode empirique qui consiste à découper l'intervalle du logarithme des valeurs observées non-nulles en 4 classes de largeur égale. Les figures 7.7 et 7.6 correspondent aux fréquences des classes de flore levée pour chaque espèce.

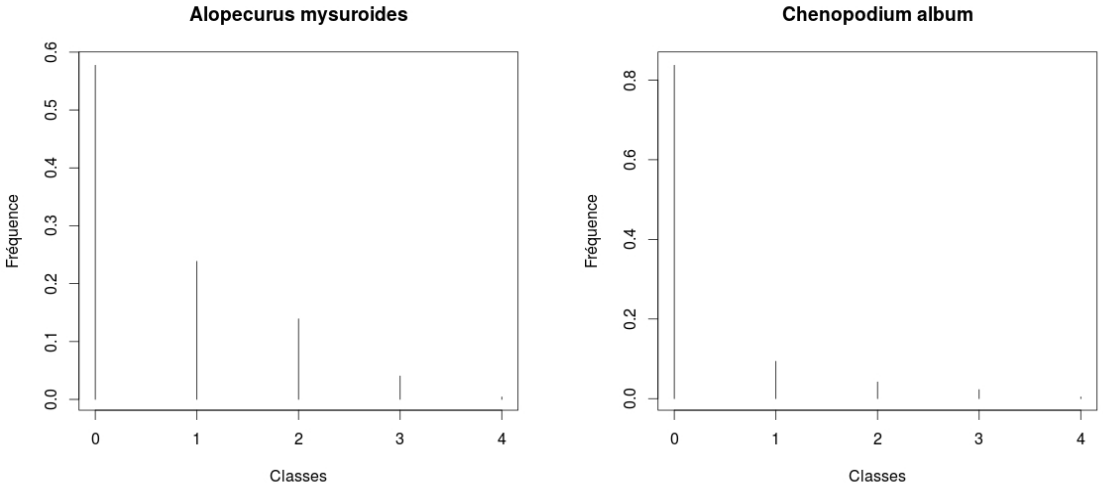


FIGURE 7.6 – Histogramme des classes des espèces

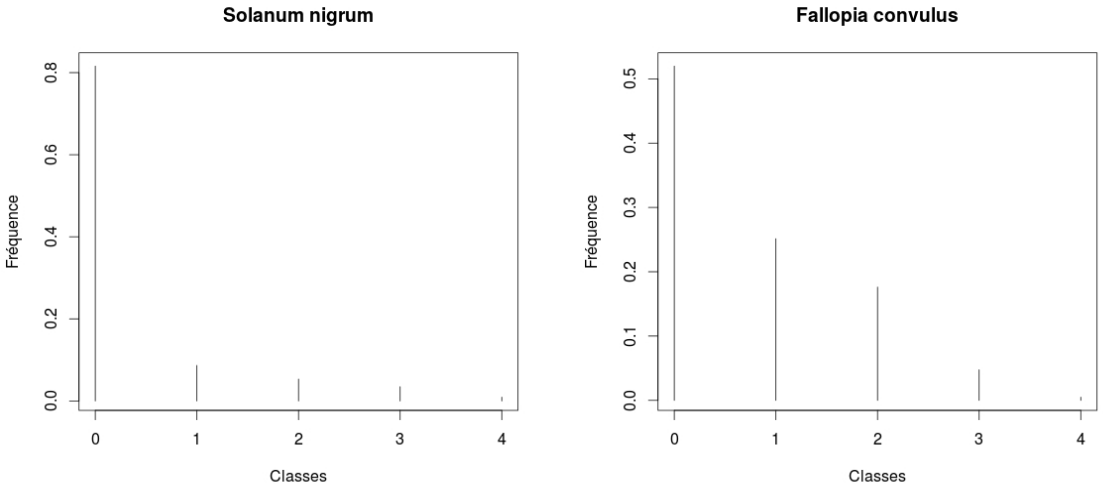


FIGURE 7.7 – Histogramme des classes des espèces

7.3 Choix de la distance de colonisation

Pour compléter la construction du modèle, il faut définir quels patchs participent à la colonisation d'un patch donné. Regardons si une connexion entre tous les patchs par colonisation est cohérente ou s'il est plus raisonnable de penser qu'au-delà d'une certaine distance la colonisation n'intervient plus. Supposons que le groupe de voisins d'un patch soit l'ensemble des patchs du complexe d'Epoisses. On constate alors que la moyenne des états observés de tous les voisins ne dépasse jamais l'état 2. De plus, l'état moyen des voisins, obtenu par agrégation moyennée 4.3.1, est souvent égal à 0 puisqu'en général la flore levée est absente. Un grand nombre de patchs diminue la variabilité de la moyenne des états. Ainsi, il y a de fortes probabilités que tous les patchs d'une année n aient la même observation pour la variable des patchs voisins $f(Y_n^{C\setminus c})$.

Ainsi on peut dire qu'une modélisation dans laquelle tous les patchs peuvent se coloniser entre eux n'est pas adéquate. On sait que la colonisation dépend de la distance entre le patch émetteur et le patch receveur. Afin d'éviter d'inclure dans un voisinage les patchs qui ne contribuent pas à la colonisation d'un patch donné, il est possible de prendre la moyenne des voisins les plus proches. Pour ce faire, il est nécessaire de déterminer la distance de colonisation à partir de laquelle la colonisation ne prend plus effet. Pour cela, nous avons analysé la moyenne de la corrélation de Bravais-Pearson entre la classe de la flore levée d'un patch au temps $n-1$ et les classes des autres patchs au temps n.

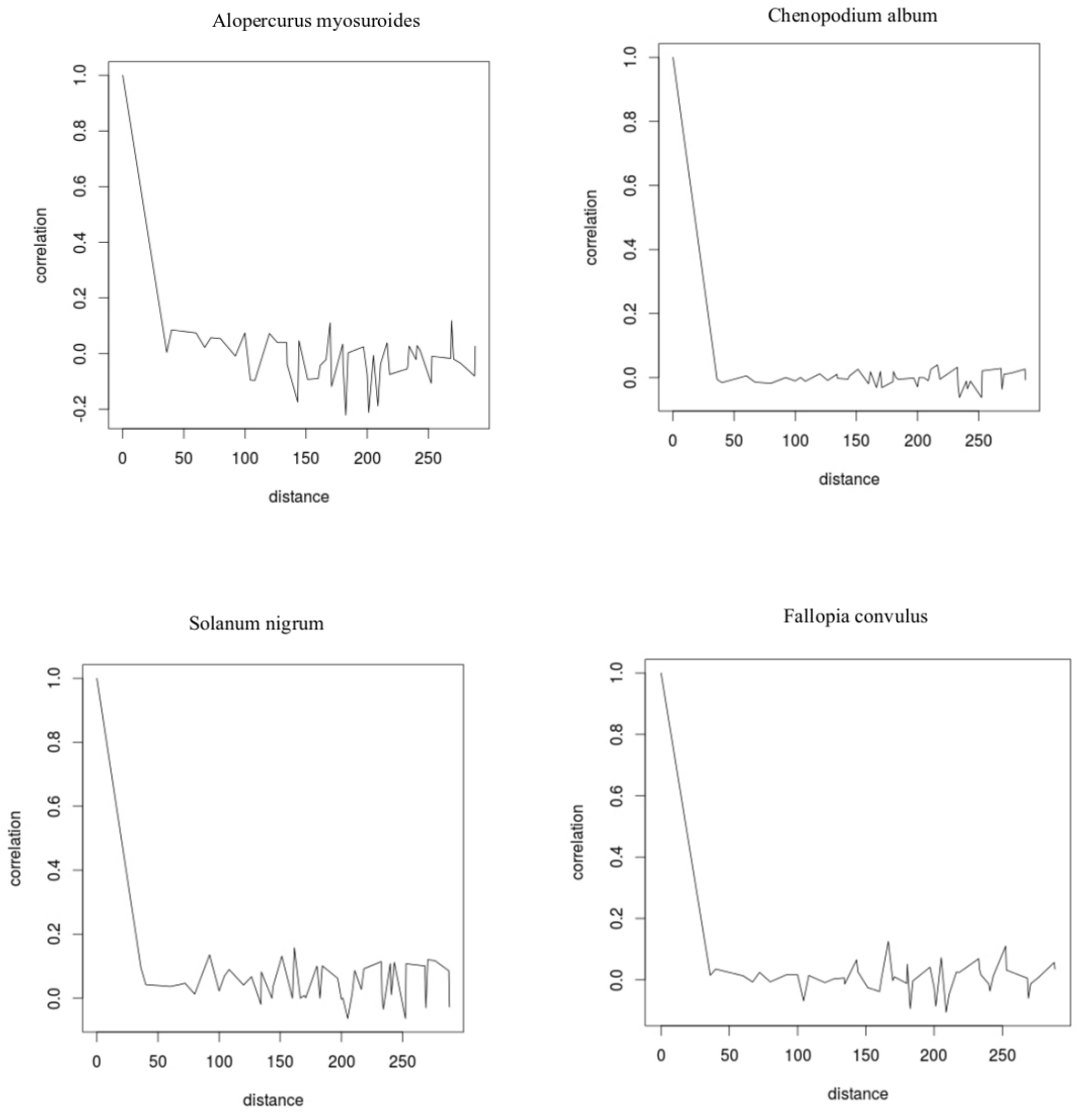


FIGURE 7.8 – Corrélation moyenne entre la classe de flore levée au temps n d'un patch et la classe de flore levée au temps $n+1$ d'un autre patch en fonction de la distance en mètres entre les 2 patchs

La figure 7.8 montrent ces corrélations moyennes en fonction de la distance entre les patchs pour les espèces *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum* et *Fallopia convolvulus*. Les espèces *Chenopodium album* et *Fallopia convolvulus* semblent avoir une très faible corrélation et la corrélation pour les espèces *Alopecurus myosuroides* et *Solanum nigrum* oscille entre 0.05 et 0.1 pour une distance inférieure à 60 mètres. On remarque qu'à partir de plus de 60 mètres, la corrélation pour toute espèce devient négligeable. Par conséquent, les voisins d'un patch sont définis comme les patchs à une distance inférieure à 60 mètres. Un patch peut donc avoir au maximum 6 voisins. Ainsi on obtient que les états maximaux des moyennes des classes des voisins à une distance inférieure à 60 mètres pour les espèces *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum*, *Fallopia convolvulus* sont respectivement $4, 4, 5, 4$.

7.4 Estimation

Dans cette section nous exposons les résultats d'estimation du MHHM-DF sans tenir compte de la culture dans un premier temps, puis en tenant compte de la culture dans la deuxième partie. Les résultats sont comparés avec ceux obtenus grâce au modèle de Pluntz et al. (2018) sur le jeu de données d'Epoisses. Dans chaque partie, on analyse les résultats du MHMM-DF puis on en extrait les probabilités de colonisation, de survie et de germination de la banque de graines. Ensuite, on réalise la prédiction de la flore levée et la restauration de la banque de graines via ce modèle.

7.4.1 Estimation sans tenir compte de la culture

7.4.1.1 Résultats de l'estimation

Les estimateurs par espèce ont été déterminés en prenant la meilleure vraisemblance parmi 8 EM. Puisque le jeu de données comporte des données manquantes à l'année 2003 pour 2 patchs, nous avons utilisé les données sur la flore levée de 88 patchs de 2000 à 2017. Les 16 premières années servent à estimer les paramètres du modèle, à partir duquel on prédit l'état de la flore levée en 2017, que l'on compare avec les données réelles. Les estimateurs sont reportés dans le tableau 7.2.

Espèces \ Estimateurs	τ	μ_1	μ_0	ν_1	ν_2	ν_3	ν_0
Alopecurus myosuroides	-0.93	3.58	-3.16	3.91	4.13	1.22	-2.22
Chenopodium album	-2.55	6.16	-4.32	8.07	-6.41	1.98	-2.80
Solanum nigrum	-1.23	9.51	-6.08	4.84	-2.18	1.53	-2.16
Fallopia convolvulus	0.21	3.77	-3.39	6.52	-1.61	0.24	-2.36
Aethusa cynapium	-0.14	5.31	-3.98	6.88	-2.56	1.43	-2.77
Galium aparine	-3.68	7.05	-5.60	5.26	-2.01	2.25	-2.08
Polygonum aviculare	-0.76	4.49	-3.32	6.85	-3.36	2.78	-2.85

TABLE 7.2 – Estimateurs du MHMM-DF sans culture pour les espèces

Pour l'*Alopecurus myosuroides*, la production de graines par la flore levée locale a le plus d'influence sur l'état de la banque de graines alors que pour les 6 autres espèces, c'est la survie de la banque de graines qui est le facteur prédominant sur l'état de la banque de graines.

En simulant 50 trajectoires à partir des estimateurs des jeux de données avec 88 patchs, on peut estimer les probabilités des états pour la banque de graines, données dans le tableau 7.3.

On remarque que si les espèces *Alopecurus myosuroides*, *Aethusa cynapium*, *Galium aparine* et *Fallopia convolvulus* visitent globalement tous les états, *Chenopodium album*, *Solanum nigrum* et *Polygonum aviculare* fréquentent souvent l'état d'extinction. On sait que si une trajectoire ne visite que des états extrêmes, en l'occurrence l'état d'exctinction, la variance des estimateurs est

	Extinction	Etat 1	Etat 2	Etat 3	Etat 4
Alopecurus myosuroides	0.26	0.24	0.17	0.15	0.18
Chenopodium album	0.58	0.25	0.10	0.05	0.02
Solanum nigrum	0.37	0.34	0.19	0.08	0.02
Fallopia convolvulus	0.21	0.19	0.17	0.20	0.24
Aethusa cynapium	0.36	0.23	0.16	0.14	0.11
Galium aparine	0.31	0.25	0.19	0.16	0.10
Polygonum aviculare	0.49	0.26	0.13	0.08	0.05

TABLE 7.3 – Estimation des probabilités des différentes classes de la banque de graines, en fonction des espèces et du jeu de données utilisé

grande. Ainsi on s'attend à avoir une forte variance pour les paramètres de la dynamique de ces 3 espèces.

7.4.1.2 Calcul de la colonisation

Afin d'évaluer l'importance de la colonisation venant des patchs voisins pour les espèces *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum* et *Fallopia convolvulus*, nous allons considérer les matrices de transition de la banque de graines quand il n'y a pas de flore levée dans le champ actuel ni aucune banque de graines au temps précédent. On peut calculer les probabilités

$$\mathbb{P}(X_{c,n} = x | Y_{c,n} = 0, X_{c,n-1} = 0, Y_n^{C\setminus c} = y^{C\setminus c})$$

Alopecurus myosuroides	Etat d'arrivée de la banque de graines				
$Y_{c,n} = 0$, $X_{c,n-1} = 0$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$f(Y_n^{\mathcal{C}\setminus c}) = 0$	0.66	0.28	0.05	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 1$	0.59	0.33	0.07	0.01	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 2$	0.52	0.37	0.10	0.01	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 3$	0.44	0.40	0.13	0.02	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 4$	0.36	0.41	0.18	0.03	0.00

TABLE 7.4 – Tableau de probabilités des états de la banque de graines en l'absence de banque de graines au temps précédent et en l'absence de flore levée pour *Alopecurus myosuroides*

Chenopodium album	Etat d'arrivée de la banque de graines				
$Y_{c,n} = 0$, $X_{c,n-1} = 0$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$f(Y_n^{\mathcal{C}\setminus c}) = 0$	0.79	0.19	0.02	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 1$	0.71	0.26	0.03	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 2$	0.61	0.32	0.07	0.01	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 3$	0.48	0.39	0.11	0.02	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 4$	0.36	0.42	0.19	0.04	0.00

TABLE 7.5 – Tableau de probabilités des états de la banque de graines en l'absence de banque de graines au temps précédent et en l'absence de flore levée pour *Chenopodium album*

Solanum nigrum	Etat d'arrivée de la banque de graines				
$Y_{c,n} = 0$, $X_{c,n-1} = 0$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$f(Y_n^{\mathcal{C}\setminus c}) = 0$	0.65	0.30	0.05	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 1$	0.56	0.35	0.08	0.01	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 2$	0.46	0.39	0.13	0.02	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 3$	0.36	0.42	0.18	0.03	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 4$	0.27	0.42	0.25	0.06	0.01

TABLE 7.6 – Tableau de probabilités des états de la banque de graines en l'absence de banque de graines au temps précédent et en l'absence de flore levée pour *Solanum nigrum*

Fallopia convolvulus	Etat d'arrivée de la banque de graines				
$Y_{c,n} = 0$, $X_{c,n-1} = 0$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$f(Y_n^{\mathcal{C}\setminus c}) = 0$	0.70	0.26	0.04	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 1$	0.69	0.27	0.04	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 2$	0.67	0.28	0.04	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 3$	0.66	0.29	0.05	0.00	0.00
$f(Y_n^{\mathcal{C}\setminus c}) = 4$	0.65	0.30	0.05	0.00	0.00

TABLE 7.7 – Tableau de probabilités des états de la banque de graines en l'absence de banque de graines au temps précédent et en l'absence de flore levée pour *Fallopia convolvulus*

La quantité $\mathbb{P}(X_{c,n} \neq 0 | Y_{c,n} = 0, X_{c,n-1} = 0, f(Y_n^{\mathcal{C}\setminus c}) = 0)$ représente la probabilité de colonisation extérieure par pluie de graines. Pour les espèces *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum*, *Fallopia convolvulus*, *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare*, la probabilité de colonisation extérieure par pluie de graines est respectivement 0.34, 0.24, 0.17, 0.33, 0.26, 0.35 et 0.24.

Comparons ces résultats avec ceux de Pluntz et al. (2018). Le modèle utilisé par Pluntz et al. (2018) est un modèle binaire avec une colonisation uniquement par pluie de graines. On peut donc considérer que cette colonisation ne distingue pas la colonisation par des patchs voisins et la colonisation extérieure à l'ensemble des patchs étudiés. Dans notre cadre du MHMM-DF adventice, la probabilité que la banque de graines soit colonisée, par l'extérieur ou par un patch voisin vaut

$$C_{moy} = 1 - \mathbb{P}(X_{c,n} = 0 | Y_{c,n} = 0, X_{c,n-1} = 0, f(Y_n^{\mathcal{C}\setminus c}) \in \{0, 1, 2, 3, 4\}).$$

En d'autres termes, c'est la probabilité que le banque de graines ne soit pas vide en l'absence de banque de graines au temps précédent et en l'absence de flore levée locale. Nous avons estimé cette quantité à l'aide de fréquences calculées par simulation. Cette même probabilité a aussi été estimée avec le modèle de Pluntz sur le jeu de données d'Epoisses. Ainsi les tableaux 7.9 7.8 regroupent ces valeurs pour le modèle de Pluntz et al. (2018) sur les données d'Epoisses et les données Biovigilance et le MHMM-DF sur les données d'Epoisses uniquement.

Données	Modèle	Espèces	C_{moy}
Biovigilance	Pluntz	*Alopecurus myosuroides*	0.09
Epoisses	Pluntz	*Alopecurus myosuroides*	0.25
Epoisses	MHMM-DF	*Alopecurus myosuroides*	0.37

TABLE 7.8 – Estimation de la probabilité de colonisation de la banque de graines pour toutes les espèces

Données	Modèle	Espèces	C_{moy}
Biovigilance	Pluntz	*Chenopodium album*	0.16
Epoisses	Pluntz	*Chenopodium album*	0.14
Epoisses	MHMM-DF	*Chenopodium album*	0.22
Biovigilance	Pluntz	*Solanum nigrum*	0.14
Epoisses	Pluntz	*Solanum nigrum*	0.13
Epoisses	MHMM-DF	*Solanum nigrum*	0.37
Biovigilance	Pluntz	*Fallopia convolvulus*	0.13
Epoisses	Pluntz	*Fallopia convolvulus*	0.19
Epoisses	MHMM-DF	*Fallopia convolvulus*	0.31
Biovigilance	Pluntz	*Aethusa cynapium*	-
Epoisses	Pluntz	*Aethusa cynapium*	0.12
Epoisses	MHMM-DF	*Aethusa cynapium*	0.24
Biovigilance	Pluntz	*Galium aparine*	0.20
Epoisses	Pluntz	*Galium aparine*	0.15
Epoisses	MHMM-DF	*Galium aparine*	0.40
Biovigilance	Pluntz	*Polygonum aviculare*	0.05
Epoisses	Pluntz	*Polygonum aviculare*	0.13
Epoisses	MHMM-DF	*Polygonum aviculare*	0.26

TABLE 7.9 – Estimation de la probabilité de colonisation de la banque de graines pour toutes les espèces

On remarque que la colonisation est plus importante dans le MHMM-DF que dans le modèle de Pluntz. De plus, les estimations obtenues pour l'*Alopecurus myosuroides* et *Polygonum aviculare* entre le jeu de données de Biovigilance et celui d'Epoisses sont très différentes.

Comme expliqué précédemment, il est possible de séparer les sources de colonisation dans le MHMM-DF. Intéressons-nous donc maintenant à la probabilité de colonisation venant uniquement des patchs voisins, notée c_{vois}. En supposant l'indépendance entre la colonisation extérieure et la colonisation voisine on pose

$$(1 - col_{ext})(1 - c_{vois}) = \mathbb{P}(X_{c,n} = 0 | Y_{c,n} = 0, X_{c,n-1} = 0, f(Y_n^{\mathcal{C}\setminus c}) \in \{1, 2, 3, 4\}),$$

où $col_{ext} = 1 - \mathbb{P}(X_{c,n} = 0 | X_{c,n-1} = 0, Y_n^C = \mathbf{0}^C)$ correspond à la probabilité de colonisation extérieure. Cette expression peut être estimée à l'aide de fréquences calculées par simulation. Le tableau 7.10 regroupe les probabilités de colonisation voisine de chaque espèce.

Espèces	col_{vois}
Alopecurus myosuroides	0.10
Chenopodium album	0.10
Solanum nigrum	0.14
Fallopia convolvulus	0.02
Aethusa cynapium	0.08
Galium aparine	0.23
Polygonum aviculare	0.12

TABLE 7.10 – Probabilités de colonisation de la banque de graines par les patchs voisins

7.4.1.3 Analyse de l'influence de la flore levée locale sur la banque de graines

Pour mesurer l'influence de la flore levée locale sur l'état de la banque de graines du patch, on peut calculer les probabilités

$$\mathbb{P}(X_{c,n} = x | Y_{c,n} = y, X_{c,n-1} = 0, Y_n^{\mathcal{C}\setminus c} = \prime^{\mathcal{C}\setminus c})$$

pour $x \in \Omega_X$ et $y \in \Omega_Y$. En effet, dans ce cas, il n'y a ni colonisation voisine, ni survie de la banque de graines donc seules la flore levée et la colonisation extérieure influent sur la banque de graines. Néanmoins, il est évidemment impossible que la banque de graines soit absente au temps $n-1$ et que la flore levée soit présente au temps n. Fort heureusement, la loi d'émission du MHMM-DF empêche que cela ne se produise dans le modèle. Mais ces probabilités sont tout de même présentes parmi les entrées de la matrice de transition et sont rassemblées dans le tableau 7.11 pour l'espèce *Alopecurus myosuroides*.

Alopecurus myosuroides	Etat d'arrivée de la banque de graines				
$Y_n^{\mathcal{C}\setminus c} = \mathbf{0}^{\mathcal{C}\setminus c}$, $X_{c,n-1} = 0$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$Y_{c,n} = 0$	0.66	0.29	0.05	0.00	0.00
$Y_{c,n} = 1$	0.41	0.41	0.15	0.03	0.00
$Y_{c,n} = 2$	0.17	0.38	0.32	0.12	0.02
$Y_{c,n} = 3$	0.04	0.19	0.36	0.31	0.10
$Y_{c,n} = 4$	0.00	0.05	0.21	0.42	0.31

TABLE 7.11 – Tableau des probabilités des états de la banque de graines en l'absence de flore levée dans le voisinage et de banque de graines au temps précédent pour *Alopecurus myosuroides*

Ainsi, pour l'*Alopecurus myosuroides*, la probabilité de production de graines non-dispersées par la flore levée locale dans l'état maximal en l'absence de banque de graines vaut 1.

7.4.1.4 Calcul de la survie des graines

Pour étudier l'influence de la banque de graines au temps $n-1$ sur l'état de la banque de graines au temps n, on va calculer les probabilités

$$\mathbb{P}(X_{c,n} = x | Y_{c,n} = 0, X_{c,n-1} = x', Y_n^{\mathbf{C}\setminus c} = \prime^{\mathcal{C}\setminus c})$$

pour x, $x' \in \Omega_X$. Cela revient à étudier la loi de la banque de graines sachant que les flores levées de tous les patchs locaux sont absentes.

Alopecurus myosuroides	Etat d'arrivée de la banque de graines				
$Y_n^C = \mathbf{0}^C$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$X_{c,n-1} = 0$	0.66	0.29	0.05	0.00	0.00
$X_{c,n-1} = 1$	0.43	0.40	0.14	0.02	0.00
$X_{c,n-1} = 2$	0.19	0.39	0.30	0.10	0.01
$X_{c,n-1} = 3$	0.05	0.22	0.37	0.28	0.08
$X_{c,n-1} = 4$	0.01	0.07	0.25	0.42	0.26

TABLE 7.12 – Tableau de probabilités des états de la banque de graines en l'absence de flore levée pour *Alopecurus myosuroides*

D'après le tableau 7.15 pour l'espèce *Fallopia convolvulus*, on a

$$\mathbb{P}(X_{c,n} = 0 | X_{c,n-1} = 3, Y_n^C = \mathbf{0}^C) = \mathbb{P}(X_{c,n} = 0 | X_{c,n-1} = 4, Y_n^C = \mathbf{0}^C) \approx 0,$$

Chenopodium album	Etat d'arrivée de la banque de graines				
$Y_n^C = \mathbf{0}^C$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$X_{c,n-1} = 0$	0.79	0.19	0.02	0.00	0.00
$X_{c,n-1} = 1$	0.35	0.42	0.19	0.04	0.00
$X_{c,n-1} = 2$	0.02	0.15	0.34	0.35	0.13
$X_{c,n-1} = 3$	0.00	0.01	0.06	0.32	0.61
$X_{c,n-1} = 4$	0.00	0.00	0.00	0.09	0.90

TABLE 7.13 – Tableau de probabilités des états de la banque de graines en l'absence de flore levée pour *Chenopodium album*

Solanum nigrum	Etat d'arrivée de la banque de graines				
$Y_n^C = \mathbf{0}^C$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$X_{c,n-1} = 0$	0.65	0.30	0.05	0.00	0.00
$X_{c,n-1} = 1$	0.35	0.42	0.19	0.04	0.00
$X_{c,n-1} = 2$	0.10	0.31	0.37	0.19	0.04
$X_{c,n-1} = 3$	0.01	0.09	0.29	0.40	0.21
$X_{c,n-1} = 4$	0.00	0.01	0.10	0.37	0.51

TABLE 7.14 – Tableau de probabilités des états de la banque de graines en l'absence de flore levée pour *Solanum nigrum*

Fallopia convolvulus	Etat d'arrivée de la banque de graines				
$Y_n^C = \mathbf{0}^C$	$X_{c,n}$= 0	$X_{c,n}$=1	$X_{c,n}$=2	$X_{c,n}$=3	$X_{c,n}$=4
$X_{c,n-1} = 0$	0.70	0.26	0.04	0.00	0.00
$X_{c,n-1} = 1$	0.30	0.42	0.22	0.05	0.00
$X_{c,n-1} = 2$	0.04	0.19	0.36	0.31	0.10
$X_{c,n-1} = 3$	0.00	0.02	0.13	0.39	0.46
$X_{c,n-1} = 4$	0.00	0.00	0.02	0.18	0.80

TABLE 7.15 – Tableau de probabilités des états de la banque de graines en l'absence de flore levée pour *Fallopia convolvulus*

ce qui signifie que si l'état de la banque de graines vaut 3 ou 4, alors elle survit forcément jusqu'à l'année suivante. De manière analogue, d'après le tableau 7.12, quand la banque de graines de l'*Alopecurus myosuroides* est dans l'état maximal, elle a une probabilité égale à 0.99 de survivre l'année suivante.

Regardons maintenant les estimateurs de la probabilité de survie de la banque de graines via le modèle de Pluntz et al. (2018). On sait que $\mathbb{P}(X_{c,n} = 0 | X_{c,n-1} \in \{1, \dots, 4\}, Y_n^C = \mathbf{0}^C)$ correspond à la probabilité d'échec simultané de la survie des graines et de la colonisation extérieure. On suppose que la colonisation extérieure est indépendante de la survie de la banque de graines. De ce fait, on a

$$(1 - col_{ext}) \times (1 - s) = \mathbb{P}(X_{c,n} = 0 | X_{c,n-1} \in \{1, \dots, 4\}, Y_n^C = \mathbf{0}^C)$$

où s correspond à la probabilité de survie de la banque de graines. A l'aide des simulations, on peut donc calculer de façon empirique la probabilité s. Le tableau 7.16 donne la probabilité de survie de la banque de graines estimée par le modèle de Pluntz et al. (2018) et par le MHMM-DF.

Afin de calculer le temps moyen avant l'extinction de la banque de graines, on utilisera la loi géométrique. La loi géométrique correspond à une succession de tirages de Bernoulli jusqu'au premier succès. Ici, le succès représente l'échec de survie de la banque de graines, de probabilité $1 - s$. Ainsi, l'espérance d'une telle loi géométrique est $1/(1 - s)$.

Données	Modèles	Espèces	s	$1/(1-s)$
Biovigilance	Pluntz	*Alopecurus myosuroides*	0.51	2.04
Epoisses	Pluntz	*Alopecurus myosuroides*	0.52	2.08
Epoisses	MHMM-DF	*Alopecurus myosuroides*	0.52	2.08
Biovigilance	Pluntz	*Chenopodium album*	0.81	5.26
Epoisses	Pluntz	*Chenopodium album*	0.82	5.55
Epoisses	MHMM-DF	*Chenopodium album*	0.65	2.86
Biovigilance	Pluntz	*Solanum nigrum*	0.89	9.1
Epoisses	Pluntz	*Solanum nigrum*	0.75	4
Epoisses	MHMM-DF	*Solanum nigrum*	0.58	2.38
Biovigilance	Pluntz	*Fallopia convolvulus*	0.92	12.5
Epoisses	Pluntz	*Fallopia convolvulus*	0.82	5.55
Epoisses	MHMM-DF	*Fallopia convolvulus*	0.75	4
Biovigilance	Pluntz	*Aethusa cynapium*	—	—
Epoisses	Pluntz	*Aethusa cynapium*	0.76	4.17
Epoisses	MHMM-DF	*Aethusa cynapium*	0.62	2.63
Biovigilance	Pluntz	*Galium aparine*	0.76	4.17
Epoisses	Pluntz	*Galium aparine*	0.81	5.26
Epoisses	MHMM-DF	*Galium aparine*	0.69	3.26
Biovigilance	Pluntz	*Polygonum aviculare*	0.91	11.11
Epoisses	Pluntz	*Polygonum aviculare*	0.69	3.23
Epoisses	MHMM-DF	*Polygonum aviculare*	0.55	2.22

TABLE 7.16 – Tableau des probabilités de survie de la banque de graines selon les modèles

Pour toutes les espèces, l'estimation de la survie par MHMM-DF est inférieure à celle par le modèle de Pluntz.

7.4.1.5 Calcul de la germination

Considérons maintenant la probabilité de présence de la flore levée sachant la banque de graines, c'est-à-dire les probabilités $\mathbb{P}(Y_{c,n} \neq 0 | X_{c,n-1} \in \{0,1,2,3,4\})$.

MHMM-DF	Probabilité de présence de flore levée par état de la banque de graines				
	$X_{c,n-1}$= 0	$X_{c,n-1}$=1	$X_{c,n-1}$=2	$X_{c,n-1}$=3	$X_{c,n-1}$=4
Alopecurus myosuroides	0	0.33	0.53	0.75	0.91
Chenopodium album	0	0.16	0.44	0.82	0.98
Solanum nigrum	0	0.06	0.32	0.88	0.99
Fallopia convolvulus	0	0.24	0.43	0.67	0.88
Aethusa cynapium	0	0.19	0.44	0.77	0.96
Galium aparine	0	0.58	0.21	0.60	0.94
Polygonum aviculare	0	0.29	0.54	0.82	0.96

TABLE 7.17 – Tableau des probabilités de présence de flore levée par état de la banque de graines

On remarque que *Solanum nigrum* a une probabilité de présence de flore levée très grande quand la banque de graines est dans un état supérieur à 2.

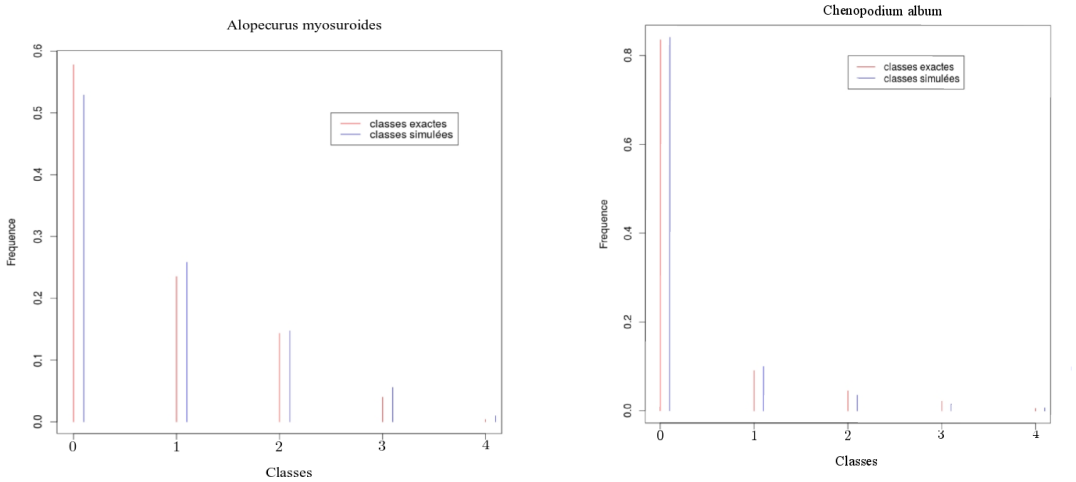
Le paramètre $g = \mathbb{P}(Y_{c,n} \neq 0 | X_{c,n-1} \neq 0)$ dans le modèle de Pluntz et al. (2018) correspond à la probabilité de germination sachant la présence de graines dans la banque de graines. Cette probabilité dans le MHMM-DF peut être calculée à l'aide de simulations. On regroupe dans le

tableau 7.18 les valeurs de $\mathbb{P}(Y_{c,n} \neq 0 | X_{c,n-1} \neq 0)$ pour le modèle de Pluntz et al. (2018) et le MHMM-DF sur les données Biovigilance et celles d'Epoisses.

Données	Modèles	Espèces	g
Biovigilance	Pluntz	*Alopecurus myosuroides*	0.59
Epoisses	Pluntz	*Alopecurus myosuroides*	0.67
Epoisses	MHMM-DF	*Alopecurus myosuroides*	0.56
Biovigilance	Pluntz	*Chenopodium album*	0.57
Epoisses	Pluntz	*Chenopodium album*	0.32
Epoisses	MHMM-DF	*Chenopodium album*	0.35
Biovigilance	Pluntz	*Solanum nigrum*	0.35
Epoisses	Pluntz	*Solanum nigrum*	0.36
Epoisses	MHMM-DF	*Solanum nigrum*	0.27
Biovigilance	Pluntz	*Fallopia convolvulus*	0.37
Epoisses	Pluntz	*Fallopia convolvulus*	0.60
Epoisses	MHMM-DF	*Fallopia convolvulus*	0.58
Biovigilance	Pluntz	*Aethusa cynapium*	—
Epoisses	Pluntz	*Aethusa cynapium*	0.56
Epoisses	MHMM-DF	*Aethusa cynapium*	0.51
Biovigilance	Pluntz	*Galium aparine*	0.56
Epoisses	Pluntz	*Galium aparine*	0.52
Epoisses	MHMM-DF	*Galium aparine*	0.35
Biovigilance	Pluntz	*Polygonum aviculare*	0.40
Epoisses	Pluntz	*Polygonum aviculare*	0.49
Epoisses	MHMM-DF	*Polygonum aviculare*	0.48

TABLE 7.18 – Probabilités de présence de flore levée lorsque la banque de graines est présente

On remarque que pour toutes les espèces, sauf pour *Galium aparine*, les probabilités de germination via le MHM-DF et le modèle de Pluntz sont similaires. Cependant, les estimations obtenues par le MHMM-DF sont légèrement inférieures pour toutes les espèces sauf pour *Chenopodium album*. On constate aussi que les estimateurs via le modèle de Pluntz sur les données Biovigilance et celles d'Epoisses sont proches, à l'exception des espèces *Chenopodium album* et *Fallopia convolvulus*.



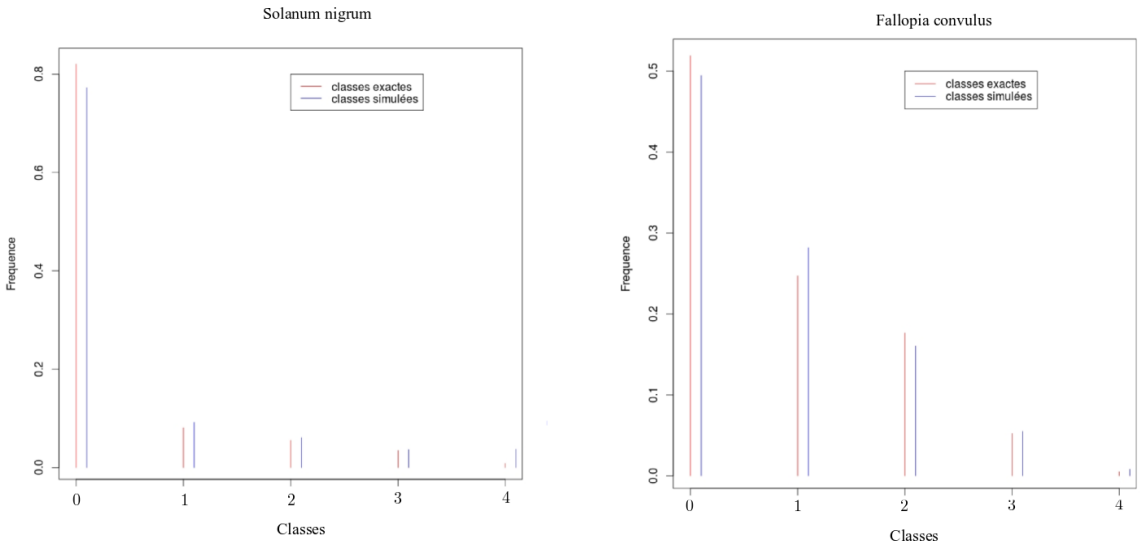


FIGURE 7.9 – Histogramme des classes exactes et simulées selon les estimateurs des espèces

Lorsqu'on compare les histogrammes des classes d'abondance des données Epoisses et ceux obtenus en simulant des trajectoires dans la figure 7.9, on remarque que, même si la distribution des classes est similaire, l'état d'extinction est sous-estimé, à part pour *Chenopodium album*. Cela implique que la probabilité de germination est surestimée dans le modèle. Les graphes pour les espèces *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare* se trouvent en annexe E.

7.4.1.6 Qualité de la prédiction de la flore levée

La qualité de la prédiction du modèle sur les données d'Epoisses est évaluée grâce aux estimateurs obtenus sur les données de 2000 à 2016 utilisés pour prédire la flore levée de l'année 2017. Les tableaux 7.19, 7.20 correspondent aux matrices de confusion des prédictions.

Alopecurus myosuroides	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	38	9	1	0	0
	1	12	7	0	0	0
	2	9	8	0	0	0
	3	1	3	0	0	0
	4	0	0	0	0	0

Solanum nigrum	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	82	0	0	0	0
	1	2	0	0	0	0
	2	4	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0

Chenopodium album	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	77	0	0	0	0
	1	0	0	0	0	0
	2	6	0	0	0	0
	3	2	0	0	0	0
	4	3	0	0	0	0

Fallopia convolvulus	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	46	8	0	0	0
	1	8	5	0	0	0
	2	8	3	0	0	0
	3	6	3	0	0	0
	4	0	1	0	0	0

TABLE 7.19 – Matrices de confusion des prédictions de la flore levée de l'année 2017

Aethusa cynapium	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	69	1	0	0	0
	1	1	0	0	0	0
	2	8	0	0	0	0
	3	6	0	0	0	0
	4	2	0	0	0	0

Polygonum aviculare	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	79	0	0	0	0
	1	1	0	0	0	0
	2	4	1	0	0	0
	3	1	1	0	0	0
	4	1	0	0	0	0

Galium aparine	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	61	0	0	0	0
	1	4	1	0	0	0
	2	17	1	0	0	0
	3	5	0	0	0	0
	4	0	0	0	0	0

TABLE 7.20 – Matrices de confusion des prédictions de la flore levée de l'année 2017

Sur les 88 prédictions pour *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum*, *Fallopia convolvulus*, *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare*, le nombre de prédictions correctes vaut respectivement 45, 77, 82, 51, 70, 62, 79 . Les prédictions sous-estiment la densité des futures flores levées. En effet, pour *Chenopodium album* et *Solanum nigrum*, le modèle prédit que la flore levée de l'année 2017 sera dans l'état d'extinction pour les 88 patchs. Les prédictions sont meilleures pour les espèces *Chenopodium album* et *Solanum nigrum* que pour les espèces *Alopecurus myosuroides* et *Fallopia convolvulus*. Plus de la moitié des prédictions par le MHMM-DF sont correctes alors qu'une prédiction au hasard ne serait correcte que dans 20% des cas.

7.4.1.7 Qualité de la restauration de la banque de graines

Le modèle MHMM-DF fournit une restauration des états de la banque de graines mais il est naturel de se demander si l'on peut associer à chaque état une fourchette à un nombre de graines par mètre carré. Même si les états de la flore levée ont été choisis comme classes d'abondance à partir des données et que l'on peut donc savoir à quel intervalle de nombres de plantes il correspond, nous n'avons aucune information sur l'échelle de différence entre les classes de la banque de graines et les classes de la flore levée.

Notre modèle suppose que l'état 0 correspond à l'extinction et que les états sont croissants, c'est-à-dire que le nombre de graines à l'état 1 est inférieur à celui dans l'état 2 et ainsi de suite.

Les données d'Epoisses fournissent le nombre de graines de la banque, via carottage, en 2000, 2001, 2005 et 2010. Des données partielles ont aussi été récoltées en 2002 et 2003. On dispose donc de 48 données par espèce pour évaluer la qualité de nos restaurations. Pour ce faire, pour chaque classe, on va calculer le minimum, le maximum et la moyenne des densités de graines des années pour lesquelles le modèle attribue cette classe à l'état de la banque de graines. Si la restauration de la banque de graines est bonne, on s'attend à avoir des moyennes, des minima et des maxima croissants en termes de classes croissantes.

Espèces / Etats de $\hat{X}$	0	1	2	3	4
Alopecurus myosuroides	0	3	4	11	30
Chenopodium album	40	5	1	1	1
Solanum nigrum	25	12	7	1	3
Fallopia convolvulus	16	15	3	5	9
Aethusa cynapium	12	8	4	14	10
Galium aparine	39	2	2	2	3
Polygonum aviculare	29	11	2	3	3

FIGURE 7.10 – Nombre d'occurrences des classes estimées parmi les 48 données disponibles

D'après le tableau 7.10, l'algorithme prédit que les graines de l'*Alopecurus myosuroides* dans la banque de graines sont très présentes, contrairement aux espèces *Chenopodium album* et *Solanum nigrum*. On dispose seulement de 48 données par espèce pour la banque de graines et de nombreux états sont restaurés moins de 5 fois, ce qui rend très sensible la détermination d'un intervalle de densité qui représenterait chaque classe.

Comme le résume le tableau 7.21, seule la restauration pour l'espèce *Alopecurus myosuroides* donne des moyennes de densité de graines croissantes avec les classes. Toutes les classes s'entrecoupent. Ainsi deux prédictions différentes de l'état de la banque de graines peuvent correspondre à la même densité de graines, ce qui n'est pas satisfaisant.

	Minimum de la densité au mètre carré quand $\hat{X} = j$				
	$\hat{X} = 0$	$\hat{X} = 1$	$\hat{X} = 2$	$\hat{X} = 3$	$\hat{X} = 4$
Alopecurus myosuroides	NA	0.00	0.00	0.00	125.86
Chenopodium album	62.93	0.00	881.02	503.44	40212.27
Solanum nigrum	62.93	0.00	0.00	1762.04	62.93
Fallopia convolvulus	0.00	62.93	125.86	62.93	0.00
Aethusa cynapium	0.00	0.00	0.00	0.00	125.86
Galium aparine	0.00	0.00	125.86	1887.90	251.72
Polygonum aviculare	0.00	0.00	755.16	125.86	503.44
	Maximum de la densité au mètre carré quand $\hat{X} = j$				
	$\hat{X} = 0$	$\hat{X} = 1$	$\hat{X} = 2$	$\hat{X} = 3$	$\hat{X} = 4$
Alopecurus myosuroides	NA	1384.46	22088.43	152982.83	611239.09
Chenopodium album	21773.78	2957.71	881.02	503.44	40212.27
Solanum nigrum	18941.93	507089.94	78662.50	1762.04	19067.79
Fallopia convolvulus	3838.73	4719.75	9250.71	25801.30	9502.43
Aethusa cynapium	22151.36	28192.64	1950.83	3272.36	1824.97
Galium aparine	4027.52	0.00	2643.06	2013.76	2202.55
Polygonum aviculare	4908.54	503.44	2013.76	7866.25	14536.83
	Moyenne de la densité au mètre carré quand $\hat{X} = j$				
	$\hat{X} = 0$	$\hat{X} = 1$	$\hat{X} = 2$	$\hat{X} = 3$	$\hat{X} = 4$
Alopecurus myosuroides	NA	755.16	5883.95	19159.32	81829.98
Chenopodium album	2098.72	1094.98	881.02	503.44	40212.27
Solanum nigrum	2444.20	58346.60	16982.11	1762.04	8076.02
Fallopia convolvulus	1183.87	1757.84	4572.91	7450.91	1643.17
Aethusa cynapium	5836.76	8078.64	692.23	1038.35	711.11
Galium aparine	288.83	0.00	1384.46	1950.83	1405.44
Polygonum aviculare	757.33	85.81	1384.46	3125.52	9502.43

TABLE 7.21 – Minima, maxima et moyennes des densités de graines par classe estimée et par espèce

7.4.2 Estimation selon la culture

On souhaite maintenant prendre en compte la saison de la culture dans nos estimations. On va distinguer deux types de dynamique, une associée aux cultures d'été et l'autre aux cultures d'hiver. Comme précédemment, nos estimations reposent sur les données d'Epoisses sur les 88 patchs de 2000 à 2016. Les 2 patchs pour lesquels les données en 2003 sont manquantes ne sont pas pris en compte. On estime deux jeux de paramètres pour chaque espèce, le premier en utilisant les données dans les cultures d'été, et le second dans celles d'hiver.

Dans un premier temps, on veut déterminer si la prise en compte de la saison de la culture dans laquelle est présente l'adventice étudiée permet d'établir un modèle plus adéquat de sa dynamique.

Pour cela, on utilise le AIC, calculé pour chacun des modèles dans le tableau 7.22.

	AIC culture	AIC sans culture
Alopecurus myosuroides	2628.46	2597.55
Chenopodium album	1586.45	1633.25
Solanum nigrum	1709.69	1839.05
Fallopia convolvulus	2997.56	3051.25
Aethusa cynapium	2736.07	2759.01
Galium aparine	2021.74	2057.93
Polygonum aviculare	2227.70	2264.24

TABLE 7.22 – Tableau des AIC des modèles avec et sans la saison de culture pour toutes les espèces

A l'exception de *Alopecurus myosuroides*, toutes les espèces de notre étude sont mieux décrites par un modèle qui tient compte la saison de la culture.

Dans un second temps, en simulant 50 trajectoires à partir des estimateurs d'été puis d'hiver obtenus avec 88 patchs sur 14 ans, on peut calculer les fréquences des états de la banque de graines, renseignées dans le tableau 7.23.

Espèces	Etats / Saisons	0	1	2	3	4
Alopecurus myosuroides	hiver	0.26	0.23	0.16	0.14	0.20
	été	0.33	0.29	0.17	0.12	0.09
Chenopodium album	hiver	0.59	0.30	0.09	0.02	0.00
	été	0.13	0.17	0.19	0.23	0.28
Solanum nigrum	hiver	0.61	0.31	0.07	0.01	0.00
	été	0.45	0.32	0.14	0.06	0.02
Fallopia convulus	hiver	0.16	0.19	0.19	0.22	0.24
	été	0.25	0.12	0.08	0.09	0.46
Aethusa cynapium	hiver	0.30	0.19	0.15	0.16	0.19
	été	0.40	0.23	0.15	0.13	0.09
Galium aparine	hiver	0.42	0.23	0.12	0.10	0.12
	été	0.19	0.26	0.27	0.20	0.08
Polygonum aviculare	hiver	0.59	0.26	0.09	0.04	0.01
	été	0.15	0.14	0.11	0.17	0.43

TABLE 7.23 – Tableau de fréquences des états simulés de la banque de graines

Cette étude révèle que les espèces *Chenopodium album*, *Solanum nigrum* et *Polygonum aviculare* fréquentent souvent l'état d'extinction en hiver. D'après les figures 6.16.3, on sait que si une trajectoire ne visite que des états extrêmes, alors la variance des estimateurs en est grande. On s'attend donc à une forte variance pour les paramètres de *Chenopodium album*, *Solanum nigrum* et *Polygonum aviculare* en hiver.

Les jeux de paramètres pour chaque espèce en fonction de la saison de culture sont présentés dans le tableau 7.24. Le paramètre dont l'influence prédomine sur la dynamique de l'adventice est surligné. On donne les probabilités de colonisation de la banque de graines, de survie et de germination fournies par les modèles étudiés jusque-là dans les tableaux 7.27, 7.26.

Paramètres / Espèces		Culture d'hiver						Culture d'été					
	τ	μ_1	μ_0	ν_1	ν_2	ν_3	ν_0	μ_1	μ_0	ν_1	ν_2	ν_3	ν_0
Alopecurus myosuroides	0.00	3.41	-2.96	3.90	4.55	0.89	-2.25	3.76	-3.33	3.78	2.87	2.47	-2.28
Chenopodium album	-2.55	8.68	-6.21	4.84	-2.58	0.69	-2.50	5.17	-3.41	7.20	-5.68	2.60	-1.83
Solanum nigrum	-1.23	10.53	-5.84	4.36	-2.19	1.32	-2.49	3.74	-1.48	3.50	0.46	2.47	-2.60
Fallopia convolvulus	-0.21	3.66	-3.27	5.01	1.07	0.40	-2.12	3.49	-3.92	6.25	7.12	7.84	-3.21
Aethusa cynapium	-0.14	5.36	-3.92	6.92	-1.94	2.03	-2.85	5.37	-4.68	7.04	-3.60	0.00	-2.75
Galium aparine	-3.68	6.19	-4.77	5.97	-1.92	4.80	-2.53	10.44	-8.49	3.20	-0.30	0.00	-1.27
Polygonum aviculare	-0.76	4.78	-3.40	5.61	-1.65	2.68	-2.88	2.80	-2.98	7.18	-3.46	4.74	-2.63

<ws />

TABLE 7.24 – Tableau des estimateurs obtenus par MHMM-DF dépendant de la saison de culture

Les paramètres qui régissent la dynamique de la banque de graines sont le paramètre de survie ν_1, le paramètre associé à la flore levée locale ν_2, celui associé à la flore levée voisine ν_3 et ν_0 associé à la colonisation extérieure. La paramètre qui a le plus d'influence sur l'état de la banque de graines semble être pour toutes les espèces celui de survie, sauf pour *Alopecurus myosuroides* sur une culture d'hiver et *Fallopia convolvulus* sur une culture d'été. Pour l'espèce *Fallopia convolvulus* dans les cultures d'été, c'est le paramètre de colonisation voisine qui a le plus d'influence et pour *Alopecurus myosuroides* dans les culture d'hiver, c'est celui de l'influence locale de la flore levée. Les tableaux 7.27, 7.26 montrent que la survie des graines est plus forte dans une culture d'été pour *Chenopodium album*, *Polygonum aviculare*, *Galium aparine* et *Aethusa cynapium*. La survie de la banque de graines est plus élevée en hiver pour le reste des espèces. La probabilité de colonisation semble être similaire dans les culture d'hiver et d'été pour toutes les espèces sauf pour *Chenopodium album*.

7.4.3 Qualité de la prédiction de la flore levée selon la saison

On utilise la même méthode que dans la section précédente pour évaluer la prédiction du modèle sur les données d'Epoisses. Les matrices de confusion sont données en annexe (voir Tableau E.1). Elles sont très similaires à celles obtenues sans prendre en compte la saison des cultures (Tableau 7.20). Sur les 88 prédictions pour *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum*, *Fallopia convolvulus*, *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare*, le nombre de prédictions correctes est respectivement 48, 77, 82, 47, 71, 58, 79, contre 45, 77, 82, 51, 70, 62, 79 précédemment. Prendre en compte la saison de la culture n'améliore pas significativement la qualité de la prédiction. Notons que dans ce cas, l'état 2 est prédit plus fréquemment mais ne coïncide pas avec l'état réel de la flore levée.

7.4.4 Qualité de la restauration de la banque de graines par culture

On a vu que dans le cas du MHMM-DF sans prise en compte de la saison de la culture, la restauration de la banque de graines n'était pas du tout satisfaisante puisque notre tentative d'attribuer une classe d'abondance à chaque état a fourni des classes qui s'intersectent toutes. Le tableau 7.11 indique pour chaque classe le nombre de d'années entre 2000 et 2016 (sauf 2003) pour lesquelles le modèle attribue cette classe à l'état de la banque de graines.

L'algorithme prédit que les graines

Etats de $\hat{X}$ / Espèces	Hiver					Eté				
	0	1	2	3	4	0	1	2	3	4
Alopecurus myosuroides	0	2	6	8	22	0	2	0	0	8
Chenopodium album	32	4	0	1	1	7	1	1	1	0
Solanum nigrum	25	8	2	2	1	7	1	1	1	0
Fallopia convolvulus	14	11	4	4	5	4	1	2	1	2
Aethusa cynapium	11	7	6	7	7	1	1	1	5	2
Galium aparine	31	3	0	2	2	9	0	0	0	1
Polygonum aviculare	24	9	2	2	1	7	0	1	0	2

FIGURE 7.11 – Nombres d'occurrences de chaque classe dans les estimations selon la saison de culture

de *Alopecurus myosuroides* sont très présentes alors que les banques de graines pour *Chenopodium album*, *Galium aparine* sont souvent vides. On calcule pour chaque classe la moyenne des densités de graines des années pour lesquelles le modèle attribue cette classe à l'état de la banque de graines (voir tableau 7.25).

Culture d'hiver	Moyenne de densité au mètre carré quand $\hat{X} = j$				
	$\hat{X} = 0$	$\hat{X} = 1$	$\hat{X} = 2$	$\hat{X} = 3$	$\hat{X} = 4$
Alopecurus myosuroides	NA	1006.88	3943.61	26328.34	66339.66
Chenopodium album	1925.26	1132.74	NA	503.44	40212.27
Solanum nigrum	5827.32	76090.24	220.255	9565.36	5097.33
Fallopia convolvulus	1276.58	1201.39	3303.83	2674.53	5097.33
Aethusa cynapium	6218.63	8261.81	943.95	845.06	566.37
Galium aparine	217.21	922.97	NA	1069.81	1982.30
Polygonum aviculare	888.89	97.89	1384.46	314.65	7866.25
Culture d'été	Moyenne de densité au mètre carré quand $\hat{X} = j$				
	$\hat{X} = 0$	$\hat{X} = 1$	$\hat{X} = 2$	$\hat{X} = 3$	$\hat{X} = 4$
Alopecurus myosuroides	NA	125.86	NA	NA	124428.34
Chenopodium album	1360.86	NA	6639.11	NA	NA
Solanum nigrum	13898.54	26115.95	1887.9	1762.04	NA
Fallopia convolvulus	2312.677	125.86	3649.94	62.93	4751.22
Aethusa cynapium	1636.18	6796.44	1950.83	994.294	975.42
Galium aparine	503.44	NA	NA	NA	2013.76
Polygonum aviculare	98.89	NA	1384.46	NA	14001.93

TABLE 7.25 – Moyennes des densités par classe estimée et par espèce selon la saison

Alopecurus myosuroides est la seule espèce qui garde un ordre de croissance entre les classes cohérent avec les données de la banque de graines. On dénote une amélioration de la restauration pour les espèces *Galium aparine* et *Polygonum aviculare* en été mais la banque de graines de *Solanum nigrum* est toujours très mal estimée. Notons aussi qu'il n'y a pas d'amélioration sur les bornes des classes estimées.

7.4.5 Tableau récapitulatif

Espèces	Données	Modèles	s	$(1/(1-s))$	C_{moy}	col_{vois}	g
Alopecurus myosuroides	Biovigilance	Pluntz	0.51	2.04	0.09	-	0.59
	Epoisses	Pluntz	0.52	2.08	0.25	-	0.67
		MHMM-DF	0.52	2.08	0.37	0.10	0.56
		MHMM-DF hiver	0.51	2.04	0.36	0.08	0.61
		MHMM-DF été	0.46	1.85	0.37	0.21	0.47
Chenopodium album	Biovigilance	Pluntz	0.81	5.26	0.16	-	0.57
	Epoisses	Pluntz	0.82	5.55	0.14	-	0.32
		MHMM-DF	0.65	2.86	0.22	0.10	0.35
		MHMM-DF hiver	0.46	1.85	0.27	0.048	0.12
		MHMM-DF été	0.82	5.55	0.56	0.42	0.74

TABLE 7.26 – Tableau des probabilités de colonisation, de survie et de germination selon les modèles

Espèces	Données	Modèles	s	$(1/(1-s))$	C_{moy}	col_{vois}	g
Solanum nigrum	Biovigilance	Pluntz	0.89	9.1	0.14	-	0.35
	Epoisses	Pluntz	0.75	4	0.13	-	0.36
		MHMM-DF	0.58	2.38	0.37	0.14	0.27
		MHMM-DF hiver	0.37	1.59	0.28	0.09	0.19
		MHMM-DF été	0.26	1.35	0.34	0.20	0.86
Fallopia convulus	Biovigilance	Pluntz	0.92	12.5	0.13	-	0.37
	Epoisses	Pluntz	0.82	5.55	0.19	-	0.60
		MHMM-DF	0.75	4	0.31	0.019	0.58
		MHMM-DF hiver	0.71	3.45	0.38	0.04	0.59
		MHMM-DF été	0.62	2.63	0.20	0.43	0.52
Aethusa cynapium	Biovigilance	Pluntz	-	-	-	-	-
	Epoisses	Pluntz	0.82	5.55	0.14	-	0.32
		MHMM-DF	0.62	2.63	0.24	0.08	0.51
		MHMM-DF hiver	0.59	2.44	0.23	0.11	0.61
		MHMM-DF été	0.69	3.23	0.22	0	0.36
Galium aparine	Biovigilance	Pluntz	0.76	4.17	0.20	-	0.56
	Epoisses	Pluntz	0.75	4	0.13	-	0.36
		MHMM-DF	0.69	3.26	0.40	0.23	0.35
		MHMM-DF hiver	0.59	2.44	0.31	0.39	0.44
		MHMM-DF été	0.73	3.70	0.63	0	0.19
Polygonum aviculare	Biovigilance	Pluntz	0.91	11.11	0.05	-	0.40
	Epoisses	Pluntz	0.82	5.55	0.19	-	0.60
		MHMM-DF	0.55	2.22	0.26	0.12	0.48
		MHMM-DF hiver	0.41	1.69	0.21	0.14	0.42
		MHMM-DF été	0.75	4	0.35	0.35	0.63

TABLE 7.27 – Tableau des probabilités de colonisation, de survie et de germination selon les modèles

7.4.6 Discussion

Les estimations obtenues par le modèle de Pluntz et al. (2018) sur les données Biovigilance et sur les données d'Epoisses sont légèrement différentes. En effet, les données Biovigilance reposent sur une étude d'environ trois ans sur 329 champs répartis dans toute la France alors que les données d'Epoisses sont récoltées sur 9 zones distantes de quelques kilomètres, de 10 champs chacune, pendant 17 ans. Dans ce cas, il est possible qu'il y ait des interactions, notamment de la colonisation, entre les populations des différents patchs. De plus, la dynamique des adventices dépend de facteurs abiotiques, comme le vent, la température et les précipitations, et biotiques, tels que la rotation de la culture et le type de désherbage utilisé. Ces facteurs sont susceptibles de différer selon le lieu de l'étude.

Comme attendu, le modèle de Pluntz estime une colonisation plus faible pour les données d'Epoisses que celle de Biovigilance pour les espèces *Chenopodium album*, *Solanum nigrum* et *Galium aparine*. Inversement, on présume que la survie est plus grande pour les données Biovigilance que pour celles d'Epoisses, comme expliqué en Annexe F. Les survies estimées vont dans ce sens, et ce pour les deux types de modèles. Les probabilités de germination estimées sont plus fortes pour les données d'Epoisses pour toutes les espèces sauf *Chenopodium album* et *Galium aparine*. La grande variabilité dans les estimations de la probabilité de germination peut être due aux différences géographiques des facteurs abiotiques.

Les estimations obtenues du modèle de Pluntz et al. (2018) et du MHMM-DF sur les données d'Epoisses diffèrent sensiblement. L'estimation de la probabilité de survie de la banque de graines

est plus grande avec le modèle de Pluntz et al. (2018). C'est aussi le cas pour la probabilité de germination sauf pour *Chenopodium album* et *Aethusa cynapium*.

Cependant, comparer les résultats obtenus sur les données d'Epoisses avec le modèle de Pluntz et al. (2018) et le MHMM-DF peut sembler non pertinent. Plusieurs éléments diffèrent entre la modélisation de Pluntz et al. (2018) et le MHMM-DF. Tout d'abord, le modèle de Pluntz et al. (2018) est un modèle avec états binaires alors que le MHMM-DF est en classes d'abondance. De plus, le modèle de Pluntz et al. (2018) ne suppose aucune interaction entre patchs et modélise la colonisation par pluie de propagules, ce qui peut paraître inapproprié car les champs dans le complexe d'Epoisses sont proches les uns des autres. Un apport notable du MHMM-DF sur le modèle de Pluntz et al. (2018) est la modélisation de la colonisation non seulement par pluie de propagules mais aussi de façon spatiale. Les colonisations dans les deux modèles sont très différentes donc il n'est pas approprié de les comparer. Enfin, Pluntz et al. (2018) fixent la probabilité que la flore levée produise des graines qui intègrent la banque de graines, appelée paramètre de reproduction, à 1. Ce paramètre a été fixé ainsi sinon le modèle n'est pas identifiable. Dans son rapport de stage, M. Pluntz prouve que dans le contexte où tous les paramètres de son modèle sont libres, l'estimation induit une corrélation négative entre les paramètres de germination et de reproduction. Fixer le paramètre de reproduction à 1 implique que la probabilité de germination peut être sous-estimée.

L'un des objectifs de ce modèle est d'établir le processus qui influence le plus la dynamique des adventices. On a prouvé que le modèle avec culture est celui qui représente le mieux les données sauf pour *Alopecurus myosuroides*. Nos résultats des estimations du MHMM-DF montrent que pour toutes les espèces étudiées la survie de la banque de graines joue le rôle le plus important dans la dynamique des adventices étudiées sauf pour l'espèce *Alopecurus myosuroides*. Rappelons cependant que d'après les simulations de trajectoires du MHMM-DF pour les espèces *Polygonum aviculare* et *Chenopodium album*, la banque de graines visite fréquemment l'état d'extinction, ce qui implique une forte variabilité dans l'estimation des paramètres de la dynamique de ces espèces. Les résultats fournis par le modèle pour ces espèces sont à donc à considérer avec précaution.

Avant de comparer nos résultats avec les informations empiriques de la littérature existante, on remarque que la probabilité de survie de la banque de graines et celle de la colonisation voisine sont estimées à l'aide de simulations en supposant indépendance mutuelle entre colonisation extérieure, colonisation voisine et survie. Toutefois ces indépendances n'ont pas été vérifiées au sein du MHMM-DF. Le taux de survie donné par BARRALIS et al. (1988) est calculé en comptant le nombre de graines non-germées qui survivent d'une année sur l'autre et représente donc une mesure expérimentale de la probabilité qu'une graine survive d'une année sur l'autre. Dans notre modèle, on entend par probabilité de survie des populations cachées la probabilité qu'au moins une graine de la banque de graines survive d'une année sur l'autre. Ces deux quantités ne sont pas comparables directement sans connaître le nombre de graines dans la banque (voir Annexe F). Néanmoins, la survie de la banque de graines est d'autant plus grande que la survie d'une graine l'est. Ainsi, le classement des espèces selon leur taux de survie devrait être le même que celui selon la probabilité de survie de la banque de graines. Le tableau 7.28 donne ce classement ainsi que la durée moyenne de vie des graines donnée dans la littérature (Arino et al., 2012) et celle estimée par le MHMM-DF.

D'après le tableau 7.28, le classement de la littérature et le classement estimé diffèrent uniquement pour *Polygonum aviculare*. La durée moyenne de vie de la banque de graines au sein du MHMM-DF est calculée comme l'espérance d'une loi géométrique ayant comme paramètre la probabilité de survie de la banque de graines. C'est donc une fonction croissante en la survie. Plus la survie est grande, plus la durée moyenne de vie est grande.

On remarque que les estimations de la durée de vie moyenne de la banque de graines sont sous-estimées par rapport aux données présentes dans la littérature.

La probabilité de germination estimée est plus forte pendant les cultures d'été pour les espèces *Polygonum aviculare*, *Chenopodium album* et *Solanum nigrum*. En revanche, la probabilité de germination estimée est plus forte pendant les cultures d'hiver pour les autres espèces. Ces résultats

Espèces	Probabilité de survie de la banque de graines [MHMM-DF]		Taux de survie d'une graine [BARRALIS et al. (1988)]		Durée de vie moyenne de la graine (années)		
					Espérance [MHMM-DF]		Littérature [Arino et al. (2012)]
	Proba	classement	Taux	classement	hiver	été	
Alopecurus myosuroides	0.52	1	0.154	2	2.08		3-4
Chenopodium album	0.65	3	0.48	4	1.85	5.55	6-8
Solanum nigrum	0.58	-	-	-	1.59	1.35	>10
Fallopia convolvulus	0.75	5	0.904	5	3.45	2.63	6-8
Aethusa cynapium	0.62	2	0.403	3	2.44	3.23	-
Galium aparine	0.69	4	0.139	1	2.44	2.70	3-4
Polygonum aviculare	0.55	-	-	-	1.69	4	6-8

TABLE 7.28 – Comparaison de la survie estimée avec la littérature

sont biologiquement cohérents, sauf pour *Fallopia convolvulus*, car d'après Arino et al. (2012) les périodes de levée des adventices *Alopecurus myosuroides*, *Chenopodium album*, *Solanum nigrum*, *Fallopia convolvulus*, *Galium aparine* et *Polygonum aviculare* sont respectivement : Automne-hiver et parfois printemps, Printemps-été, Fin printemps-début été, Printemps-été , Automne-hiver et parfois printemps, Printemps-été.

On constate un écart important d'ordre de grandeur entre les probabilités de colonisations voisine et extérieure si l'on tient compte ou non de la saison de culture. Le processus de grenaison est pris en compte dans le modèle à travers la production de la flore levée locale et la colonisation voisine. Pour *Chenopodium album* et *Fallopia convolvulus*, la colonisation voisine est dix fois supérieure en été qu'en hiver, et ceci coïncide avec leur période de grenaison qui est fin d'été d'après (Arino et al., 2012). La colonisation voisine vaut 0.39 en hiver contre 0 en été pour l'espèce *Galium aparine*, dont la période de grenaison maximale est en milieu d'automne (Taylor, 1999). La période de grenaison de chaque espèce coïncide donc avec la saison où la colonisation voisine est la plus élevée.

L'influence de la flore levée sur l'état de la banque de graines est négative pour toutes les espèces sauf pour *Alopecurus myosuroides*. Ce résultat est biologiquement incorrect car les plantes adultes du champ sont la principale source de nouvelles graines entrant dans la banque de graines. Dans notre modèle, l'état de la banque de graines n'est pas actualisé après la germination de certaines graines. Une fois qu'une graine germe, elle fait partie de la population observable et doit être décomptée de la population cachée, ce que le modèle ne fait pas. Dans la pratique, seules les graines qui n'ont pas germé sont à prendre en compte dans le processus de survie qui déterminera en partie l'état de la banque de graines au temps suivant. Cependant, le MHMM-DF ne prend pas en compte la proportion de graines qui germent et considère que la totalité des graines de la banque de graines influence l'état de la banque de graines au temps suivant, quel que soit le nombre de graines ayant germé. Ceci est sûrement la raison pour laquelle la contribution de la production de graines par la flore levée sur l'état de la banque de graines est négative pour toutes les espèces sauf pour *Alopecurus myosuroides* et pour *Fallopia convolvulus*.

La prédiction de la flore levée de l'année 2017 a obtenu des scores supérieurs à 50% pour *Alopecurus myosuroides* et *Fallopia convolvulus* et des scores supérieurs à 87% pour *Solanum nigrum* et *Chenopodium album*. Une prédiction au hasard aurait 20% des prédictions correctes. Ainsi, nos prédictions sont plutôt bonnes. Cependant, elles sous-estiment l'état de la flore levée, sauf pour *Alopecurus mysoroides* et *Fallopa convulus*.

La nature des données rend la restauration de la banque de graines difficile. Les relevés de la banque de graines sont faits par carottage sur une petite surface du patch. Il est donc possible d'avoir des données réelles montrant l'extinction de la banque de graines alors que le patch contient

des graines dans la banque de graines. Il est évident que la flore levée n'est présente que s'il y a des graines dans la banque de graines. Notre modèle ne peut pas prédire l'extinction de la banque de graines alors que l'on observe de la flore levée. Si les données nous indiquent qu'un patch ne présente pas de graines et que de la flore levée est présente sur ce patch l'année suivante, alors le carottage prélevé pour le relevé des données pour la banque de graines n'est pas représentatif de la population cachée. C'est le cas pour *Alopecurus myosuroides* dont la banque de graines de 6 patchs en 2000 ne présente pas de graines alors que les données de flore levée montrent la présence de flore adventice. Cela induit une restauration assez mauvaise de l'état de la banque de graines. Par exemple, l'état d'extinction de la banque de graines n'est jamais restauré pour *Alopecurus myosuroides*. D'après les données d'Epoisses, *Chenopodium album* et *Solanum nigrum* ont beaucoup de graines dans leur banque de graines. Cependant, les estimateurs associés ont tendance à restaurer la banque de graines comme éteinte. La restauration de la banque de graines n'est que très peu cohérente avec les données. L'algorithme de restauration n'étant pas à mettre en cause, le problème proviendrait de la modélisation choisie. L'une des façons de mieux modéliser la survie de la banque de graines serait d'utiliser un modèle Poisson binomial pour les adventices 4.5.2.2. L'identifiabilité générique d'un tel modèle serait prouvée de manière identique à celle du modèle binomial logistique (paragraphe 4.5.2.1).

Chapitre 8

Conclusion

Nous avons développé un modèle, appelé MHMM-DF, qui permet d'étudier la dynamique locale et régionale d'espèces avec stade caché à partir de données en classes d'abondance. Pour ce faire, nous avons construit le graphe du MHMM-DF dont les arêtes permettent de représenter les processus et interactions entre les variables observables et celles cachées qui régissent la dynamique étudiée. Dans ce modèle, on suppose que les interactions entre populations sur deux patchs différents ne peuvent avoir lieu que de la population observable à d'autres populations. Les interactions entre chaînes sont modélisées en agrégeant les états des variables observées voisines à l'aide d'une moyenne ou bien de l'agrégation alphabétique.

Une façon plus fine de modéliser les interactions entre chaînes aurait été d'attribuer autant de paramètres qu'il y a d'état observable où un paramètre est associé au nombre de variables observées dans un état (Gyllenberg et al., 1997). Par exemple, si la population observable d'un patch voisin est dense, on considérerait sa contribution à la colonisation plus importante que celle d'un patch dont la population observable est moindre. Néanmoins, nous n'avons pas réussi à montrer l'identifiabilité générique d'un tel modèle. La distance entre deux patchs, limitant souvent l'influence des interactions entre les populations, peut être modélisée de plusieurs manières au sein du MHMM-DF. On peut, choisir un paramètre par distance ce qui accroîtrait amplement le nombre de paramètres. Il est aussi possible d'utiliser un paramètre associé à une fonction de distance. Cependant cette méthode requiert de définir la fonction de distance au préalable. Dans le cadre des données d'Eppoises, nous avons défini une distance à partir de laquelle les patchs ayant une distance inférieure à celle-ci interagissent entre elles à l'aide de l'agrégation moyennée. Les deux premières méthodes rentrent dans le cadre d'autocorrélation spatiale en distance et la dernière correspond à de l'autocorrélation spatiale en voisinage. La comparaison de ces deux types d'autocorrélation spatiale a déjà été étudiée par Saas and Gosselin (2014).

Les interactions entre chaînes, partant de variables observées, ont été l'obstacle majeur à surmonter lors de l'estimation du modèle. L'estimation du MHMM-DF repose sur l'algorithme EM, qui se décompose en deux étapes, l'étape E et l'étape M. J'ai montré que le Forward-Backward, utilisé dans l'étape E, a une complexité algorithmique $O(N|\Omega_X|^2C)$, donc linéaire en le nombre de patchs C, et ce indépendamment de la paramétrisation du modèle. Cette complexité linéaire en le nombre de chaînes ne peut être conservée si l'on autorise les interactions entre chaînes partant de variables cachées. À l'inverse de l'étape E, la complexité algorithmique de l'étape M dépend de la paramétrisation et dans un cas non paramétrique a une complexité algorithmique identique à celle que l'étape E. Nous avons développé plusieurs façons de paramétrer le MHMM-DF afin de réduire le nombre de paramètres à estimer. Parmi les différentes façons de paramétrer les lois du MHMM-DF, nous avons prouvé que certaines paramétrisations sont génériquement identifiables. L'utilisation de loi paramétrées au sein du MHMM-DF permet de réduire le nombre de données nécessaires pour l'estimation.

L'estimation du MHMM-DF ne peut être faite sur des jeux données avec données manquantes. Les données manquantes sont un problème souvent rencontré en analyse de données, que ce soit pour des raisons financières, à cause d'erreurs humaines ou par manque de rigueur dans la récupération des données. Pour cela, plusieurs méthodes ont été développées afin de traiter les données manquantes. Deux méthodes classiques sont l'analyse sans complétion et l'imputation de données. Un exemple d'imputation de données est la méthode des plus proches voisins : on approche les données manquantes par la moyenne des données sur des voisins à une distance choisie. L'idée de l'analyse sans complétion est d'écarter les données manquantes dans un premier temps pour estimer les paramètres du modèle puis de prédire les données manquantes. Ensuite, on se sert de ces prédictions pour réestimer la dynamique du modèle. Une autre approche consisterait à considérer nos données manquantes comme des variables cachées. Cela impliquerait de réajuster l'étape E du EM au détriment de sa complexité algorithmique.

Dans notre cadre, on a choisi le nombre d'états des variables cachées égal au nombre de ceux observables par commodité. Cependant, il serait possible de ne pas fixer ce nombre et de le déterminer par maximum de vraisemblance.

Dans un modèle markovien classique, la durée de séjour d'une variable cachée dans un état n'est pas flexible et suit forcément une loi géométrique (Yu (2010), Ferguson (1980)). Une modélisation par une loi géométrique du temps de sortie d'un état n'est appropriée que si le phénomène peut être modélisé par une répétition d'expériences de Bernoulli, indépendantes et de même probabilité de succès. Par conséquent, la loi géométrique peut ne pas être adéquate afin de représenter la durée de séjour d'une variable cachée dans un état. Les modèles semi-markoviens sont une des alternatives au HMM car ils permettent une plus grande famille de distributions pour la sortie d'un état cachée Bulla (2006) Barbu and Limnios (2009). Par exemple les modèles semi-markoviens ont été utilisés comme alternative pour le traitement et la modélisation de la parole (Yu (2010)).

Dans le cadre des données d'Epoisses sur les adventices, la durée de vie moyenne est calculée comme l'espérance d'une loi géométrique de paramètre la survie de la banque de graines. Les estimations de la durée de vie moyenne de la banque de graines sont assez éloignées de celles présentes dans la littérature Arino et al. (2012). L'utilisation de modèle semi-markovien peut permettre d'améliorer la qualité de nos estimations et prédictions tout en intégrant l'âge de la banque de graines au sein du modèle.

L'un des avantages majeurs de notre étude est l'analyse simultanée de la contribution de tous les processus de la dynamique de l'espèce alors qu'en laboratoire, les paramètres qui déterminent la dynamique des populations cachées, à savoir pour les adventices la survie des graines, la colonisation ou encore la germination, sont étudiés séparément les uns des autres. Les résultats sur les données d'Epoisse coïncident avec la littérature existante. En effet, l'ordre de grandeur des paramètres de colonisation voisine en fonction de la saison de la culture est cohérent avec la période de grenaison. De même, l'ordre de grandeur des paramètres de germination selon la saison coïncide avec la période de levée des adventices. Notre modèle met en évidence la survie de la banque de graines, dont l'estimation semble cohérente avec la littérature, comme le paramètre qui prédomine dans la dynamique de la plupart des adventices étudiées. Adams et al. (2005) ont montré un phénomène similaire pour les plantes en zones humides. La prédiction de l'état de la flore levée est correcte à 50% alors qu'une prédiction aléatoire garantit 20% pour 5 classes. Elle est cependant souvent sous-estimée et des améliorations sont sûrement possibles.

Le MHMM-DF ne prend pas en compte la proportion de graines qui germent et considère que la totalité des graines de la banque de graines influence l'état de la banque de graines au temps suivant, quel que soit le nombre de graines ayant germé. Une solution envisageable à ce problème serait d'adapter des idées de Levin et al. (1984), c'est-à-dire de considérer la proportion de graines qui germent au sein du modèle. Ainsi, le processus de survie des graines dans la banque de graines ne dépendrait que du nombre de graines n'ayant pas germé et celui de germination dépendrait du nombre de graines ayant germé. Cette méthode peut être appliquée à un modèle avec classes d'abondance de façon analogue. Cependant, l'ajout de ce paramètre rend la preuve de

l'identifiabilité générique du modèle associé difficile et nous soupçonnons ce modèle de ne pas être génériquement identifiable.

L'un des autres facteurs qui influencent la dynamique des adventices est la stratégie de gestion employée dans le champ. Même si les stratégies de gestion peuvent théoriquement être prises en compte dans le MHMM-DF et qu'elles sont fournies dans les données d'Epoisses, ceci impliquerait une augmentation du nombre de paramètres dans le modèle et il faudrait plus de données pour éviter une forte variance des estimateurs. De la même façon, toutes les cultures peuvent être prises en compte dans le MHMM-DF au détriment d'une forte variance des estimateurs. Afin d'éviter cette forte variance nous avons regroupé les cultures en deux groupes les cultures d'hiver et les cultures d'été. Il est important de prendre en compte que la colonisation au sein du MHMM-DF en fonction de la saison de la culture correspond à une colonisation qui dépend de la saison de la culture du patch d'arrivée et non du patch de départ. Afin d'avoir une colonisation dépendant de la saison de la culture du patch de départ, il faudrait inclure un paramètre de colonisation pour chaque saison de culture.

Plusieurs modélisations de phénomènes écologiques, dépendent de l'échelle de l'étude. Levey et al. (2008) ont modélisé la dispersion des graines sur des longues distances pour examiner l'influence de la forme et la taille des patchs. D'après Hurlbert and Jetz (2007), l'indice de biodiversité est aussi dépendant de l'échelle de l'étude et Turner et al. (1989) ont montré qu'il est d'autant plus grand que l'échelle de l'étude est petite. D'après A. Cushman and McGarigal (2004) et Orians and Wittenberger (1991), la relation entre habitat et espèce est conditionnée par l'échelle. Une partie de la communauté scientifique explore la relation entre espèce et environnement en fonction de l'échelle de l'étude A. Cushman and McGarigal (2004) ; Kotliar and Wiens (1990) ; Allen and Hoekstra (1991). Se soulève donc la question de la nature de l'effet de l'échelle de grandeur des patchs sur l'estimation de la dynamique d'adventices. Tout d'abord, on peut remarquer qu'une variation de l'échelle des patchs induit une variation de l'abondance dans chaque patch. En revanche, plus l'échelle des patchs est petite, plus apparente sera la distribution de l'espèce (Gaston, 1996). En ce qui concerne le MHMM-DF, puisque de différentes échelles induisent des abondances différentes, il semble raisonnable de penser que les résultats obtenus en seraient aussi différents. La survie de la banque de graines et la germination dans le modèle de Pluntz et al. (2018) dépendent de la taille de chaque parcelle. La preuve se trouve dans l'annexe F. Puisque le MHMM-DF est une extension du modèle de Pluntz et al. (2018), on s'attend à un phénomène similaire.

Le Multi Hidden Markov Model with Datafeed back peut-être appliqué à plusieurs types d'espèces avec stades cachés. De plus, l'utilisation du MHMM-DF ne se limite pas à des fins d'estimation. Des groupes d'espèces peuvent être déterminés à partir des estimateurs des espèces à l'aide de méthodes de classification. Le modèle permet de simuler les effets des différentes rotations de culture sur la dynamique d'une adventice.

Chapitre 9

Notations

Notation	Définition mathématique	Interprétation dans le cas adventices
x	Élément de l'espace d'états Ω_X	Valeur de l'abondance de la banque de graines
y	Élément de l'espace d'états Ω_Y.	Valeur de l'abondance de la flore levée.
c	Indice de la chaîne $c \in \mathcal{C}$.	Indice du patch $c \in \mathcal{C}$.
n	Indice temporel $n \in \{1, ..., N\}$.	
$X_{c,n}$	Variable aléatoire cachée à valeurs dans Ω_X.	Banque de graines de l'année n du patch c.
$Y_{c,n}$	Variable aléatoire observée à valeurs dans Ω_Y.	Flore levée de l'année n du patch c.
$X_n^C = (X_{1,n}, ..., X_{C,n})$	Vecteur de variables cachées de toutes les chaînes à valeurs dans Ω_X^C.	Banques de graines de tous les patchs de l'année n.
$Y_{c,n}$	Variable aléatoire observée à valeurs dans Ω_Y.	Flore levée de l'année n du patch c.
$X_n^C = (X_{1,n}, ..., X_{C,n})$	Vecteur de variables cachées de toutes les chaînes à valeurs dans Ω_X^C.	Banques de graines de tous les patchs de l'année n.
$Y_n^C = (Y_{1,n}, ..., Y_{C,n})$	Vecteur de variables observées de toutes les chaînes à valeurs dans Ω_Y^C.	Flores levées de tous les patchs de l'année n.
$Y_n^{\mathcal{C}\setminus c} = (Y_{1,n}, \ldots, Y_{c-1,n}, Y_{c+1,n}, \ldots, Y_{C,n})$	Vecteur de toutes les variables observées de l'année n sauf celle de la chaîne c à valeurs dans $\Omega_Y^{(C-1)\times n}$.	Flores levées de tous les patchs de l'année n sauf pour le patch c.
$X^{C,N} = (X_{1,1}, ..., X_{C,N})$	Vecteur de toutes les variables cachées à valeurs dans $\Omega_X^{C\times N}$.	Banques de graines de tous les patchs pour toutes les années.
$Y^{C,N} = (Y_{1,1}, ..., Y_{C,N})$	Vecteur de toutes les variables observées à valeurs dans $\Omega_Y^{C\times N}$.	Flores levées de tous les patchs pour toutes les années.
ϕ	$\mathbb{P}(Y_{c,n} = y_n \vert X_{c,n-1} = x_{n-1})$	Matrice de germination.
A	$\mathbb{P}(X_{c,n} = x_n \vert X_{c,n-1} = x_{n-1}, Y_n^C = y_n^C)$	Probabilité d'avoir une abondance x_n de graines au temps n sachant toutes les flores levées et la banque de graines au temps précédent.
π	$\mathbb{P}(X_{c,0} = x)$ pour MHMM-DF adventices ou $\mathbb{P}(X_{c,0} = x \vert Y_{c,0} = y)$ pour MHMM-DF complet	Probabilité d'avoir une abondance x de graines au temps 0.
ζ	$\mathbb{P}(Y_{c,0} = y)$	Probabilité d'avoir une abondance y de flore levée 0.
$\lambda = (\phi, \pi, A)$	Vecteur de paramètres	Vecteur représentant la dynamique d'une adventice.

Annexe A

Résultats du EM et de l'algorithme de Gibbs sur HMM

Pour chaque simulation une initialisation des paramètres a été utilisée pour l'estimation de l'algorithme EM. Pour l'algorithme de Gibbs les 1000 dernières itérations ont été utilisées afin de calculer les probabilités d'émission et de transition. On fixe $\pi = (1/2, 1/2)$.

Voici les résultats

Modèle		$A = \begin{pmatrix} 0.875 & 0.125 \\ 0.2 & 0.8 \end{pmatrix}, \phi = \begin{pmatrix} 0.667 & 0.222 & 0.111 \\ 0.125 & 0.25 & 0.625 \end{pmatrix}$
Simulation 1	Vraisemblance modèle exact	$2.094293 * 10^{-35}$
EM	Estimations	$\hat{A} = \begin{pmatrix} 0.931 & 0.068 \\ 0.103 & 0.897 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.375 & 0.418 & 0.207 \\ 3.61 * 10^{-5} & 0.169 & 0.83 \end{pmatrix}$
	Vraisemblance EM	$1.209948 * 10^{-32}$
	BIC	-99.39707
MCMC	Estimations	$\hat{A} = \begin{pmatrix} 0.648 & 0.352 \\ 0.331 & 0.669 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.372 & 0.625 & 0.002 \\ 0.019 & 0.006 & 0.974 \end{pmatrix}$
	Vraisemblance MCMC	$1.758647 * 10^{-33}$
Simulation 2	Vraisemblance modèle exact	$2.003995 * 10^{-34}$
EM	Estimations	$\hat{A} = \begin{pmatrix} 0.9998 & 0.0002 \\ 0.027 & 0.973 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.659 & 0.228 & 0.113 \\ 0.221 & 0.357 & 0.421 \end{pmatrix}$
	Vraisemblance EM	$1.382816 * 10^{-33}$
	BIC	-101.5661
MCMC	Estimations	$\hat{A} = \begin{pmatrix} 0.616 & 0.384 \\ 0.429 & 0.571 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.454 & 0.332 & 0.213 \\ 0.649 & 0.329 & 0.022 \end{pmatrix}$
	Vraisemblance MCMC	$6.143561 * 10^{-38}$
Simulation 3	Vraisemblance modèle exact	$1.757897 * 10^{-33}$
EM	Estimations	$\hat{A} = \begin{pmatrix} 0.744 & 0.256 \\ 0.303 & 0.697 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.813 & 0.187 & 4.33 * 10^{-8} \\ 0.013 & 0.186 & 0.801 \end{pmatrix}$
	Vraisemblance EM	$2.113259 * 10^{-32}$
	BIC	-98.83942
MCMC	Estimations	$\hat{A} = \begin{pmatrix} 0.668 & 0.332 \\ 0.301 & 0.699 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.749 & 0.233 & 0.0173 \\ 0.016 & 0.211 & 0.773 \end{pmatrix}$
	Vraisemblance MCMC	$7.236794 * 10^{-33}$

Modèle		$A = \begin{pmatrix} 0.875 & 0.125 \\ 0.2 & 0.8 \end{pmatrix}, \phi = \begin{pmatrix} 0.444 & 0.333 & 0.222 \\ 0.222 & 0.333 & 0.444 \end{pmatrix}$
Simulation 1	Vraisemblance modèle exact	$3.628575 * 10^{-36}$
EM	Estimations	$\hat{A} = \begin{pmatrix} 1 & 1.53 * 10^{-9} \\ 0.404 & 0.596 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.413 & 0.271 & 0.316 \\ 5.42 * 10^{-74} & 0.999 & 2.41 * 10^{-5} \end{pmatrix}$
	Vraisemblance EM	$2.036358 * 10^{-35}$
	BIC	-105.7842
MCMC	Estimations	$\hat{A} = \begin{pmatrix} 0.43 & 0.57 \\ 0.392 & 0.608 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.389 & 0.266 & 0.345 \\ 0.397 & 0.268 & 0.334 \end{pmatrix}$
	Vraisemblance MCMC	$2.778684 * 10^{-36}$
Simulation 2	Vraisemblance modèle exact	$5.457271 * 10^{-36}$
EM	Estimations	$\hat{A} = \begin{pmatrix} 0.999 & 1.81 * 10^{-5} \\ 0.43 & 0.57 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.411 & 0.342 & 0.247 \\ 1.17 * 10^{-13} & 1 & 9.06 * 10^{-49} \end{pmatrix}$
	Vraisemblance EM	$3.132846 * 10^{-35}$
	BIC	-105.3535
MCMC	Estimations	$\hat{A} = \begin{pmatrix} 0.362 & 0.638 \\ 0.558 & 0.442 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.37 & 0.352 & 0.277 \\ 0.388 & 0.356 & 0.255 \end{pmatrix}$
	Vraisemblance MCMC	$7.348718 * 10^{-36}$
Simulation 3	Vraisemblance modèle exact	$1.423722 * 10^{-36}$
EM	Estimations	$\hat{A} = \begin{pmatrix} 1.31 * 10^{-8} & 1 \\ 0.999 & 1.05 * 10^{-6} \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.378 & 0.351 & 0.27 \\ 0.132 & 0.5 & 0.368 \end{pmatrix}$
	Vraisemblance EM	$1.017774 * 10^{-34}$
	BIC	-103.676
MCMC	Estimations	$\hat{A} = \begin{pmatrix} 0.292 & 0.708 \\ 0.281 & 0.719 \end{pmatrix}, \hat{\phi} = \begin{pmatrix} 0.266 & 0.474 & 0.26 \\ 0.267 & 0.442 & 0.291 \end{pmatrix}$
	Vraisemblance MCMC	$6.811329 * 10^{-36}$

Annexe B

Disque de colonisation

On se place dans un contexte de modélisation de la dynamique des adventices. Ainsi les processus ayant un effet sur la banque de graines sont la colonisation extérieure, la colonisation venant des voisins, la survie des graines dans le sol, et les graines non dispersées provenant de la flore levée locale.

On se fixe une parcelle c. On veut faire des groupes de colonisation parmi les autres parcelles selon la distance à la parcelle c. Fixons $0 = r_0 < r_1 < r_2 < \cdots < r_m$ des réels positifs. Pour tout $i \in \{1, m\}$, on pose $D_{c,i}$ l'ensemble des parcelles c' à une distance de c inférieure à r_i mais strictement supérieure à r_{i-1}. Soit $Y_{n+1}^{c,i} = \{Y_{l,n+1}, l \in D_{c,i}\}$ et $y_{n+1}^{c,i} \in \Omega_Y^{\#D_{c,i}}$ et

$$GD_{c,i,n+1} = \frac{1}{\#D_{c,i}} \sum_{l \in D_{c,i}} Y_{l,n+1}^{c,i}$$

qui représente la flore levée moyenne du disque $D_{c,i}$ au temps $n+1$. Posons

$$GD_{c,n+1} = (GD_{c,1,n+1}, GD_{c,2,n+1}, \ldots, GD_{c,m,n+1})$$

et $d_{c,i}$ la moyenne des distances entre le champs c et les champs dans le disque $D_{c,i}$.

La loi du sous graphe du MHMM-DF va être décrite à partir des probabilités de transition suivantes : $\mathbb{P}(Y_{c,n+1} = y_{c,n+1} | X_{c,n} = x_{c,n})$ et $\mathbb{P}(X_{c,n+1} = x_{c,n+1} | Y_{n+1}^C = y_{n+1}^C, X_{c,n} = x_{c,n})$. Une simplification peut être faite : $\mathbb{P}(X_{c,n+1} = x_{c,n+1} | Y_{n+1}^C = y_{n+1}^C, X_{c,n} = x_{c,n})$

$$\begin{aligned}
&= \sum_{gd \in \Omega(GD_{c,n+1})} \mathbb{P}(X_{c,n+1} = x_{c,n+1} | GD_{c,n+1} = gd, X_{c,n} = x_{c,n}, Y_{c,n+1} = y_{c,n+1}) \\
&\quad \times \prod_{i=1}^{m} \mathbb{P}(GD_{c,i,n+1} = gd_i | Y_{n+1}^{c,i} = y_{n+1}^{c,i}) \\
&= \mathbb{P}(X_{c,n+1} = x_{c,n+1} | GD_{c,n+1} = gd, X_{c,n} = x_{c,n}, Y_{c,n+1} = y_{c,n+1})
\end{aligned}$$

car

$$\mathbb{P}(GD_{c,i,n+1} = gd_i | Y_{n+1}^r = y_{n+1}^{c,i}) = \begin{cases} 1 & \text{si } gd_i = \frac{1}{\#D_{c,i}} \sum_{l \in D_{c,i}} y_{c,n+1} \\ 0 & \text{sinon.} \end{cases}$$

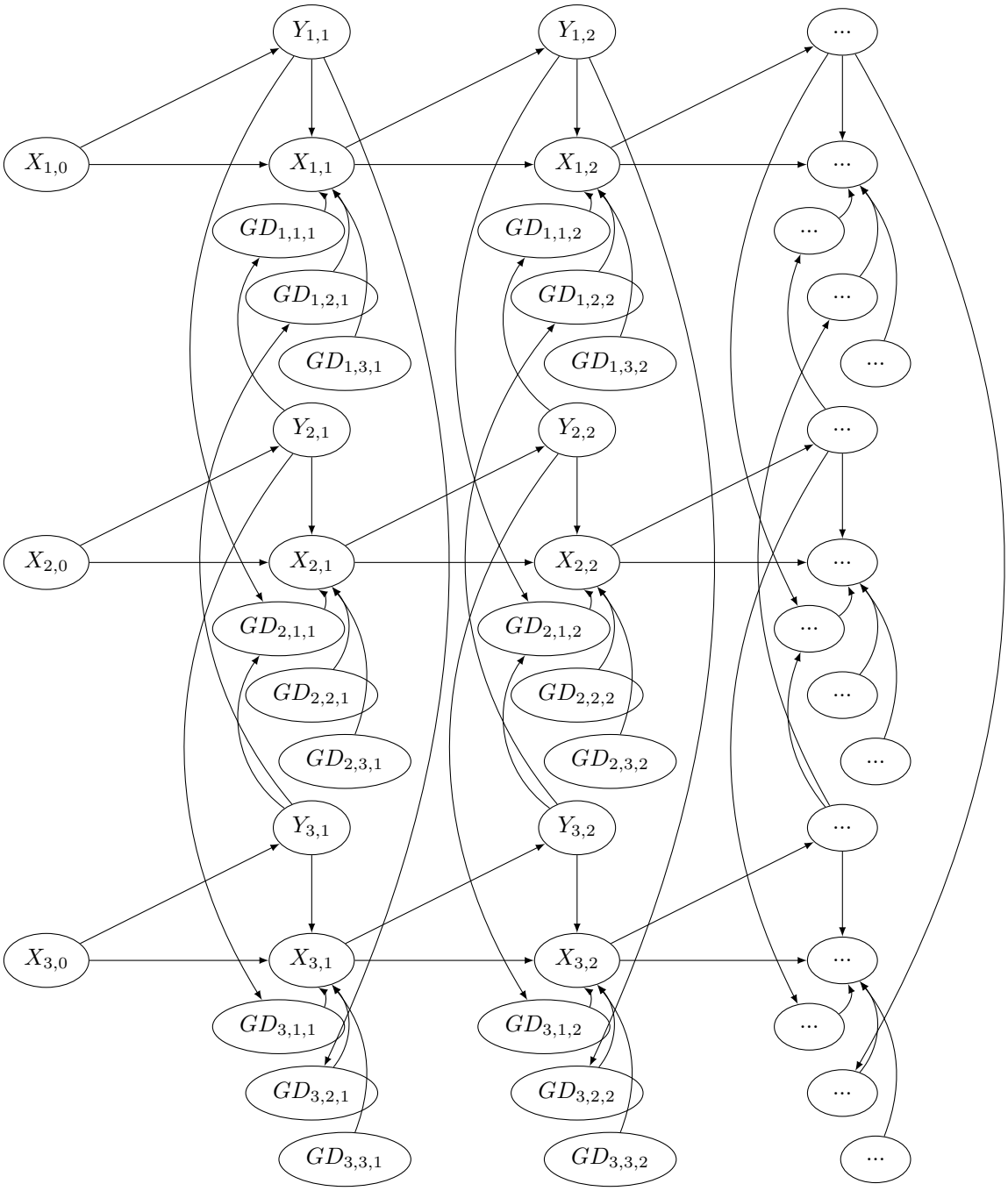


FIGURE B.1 – Le graphe du modèle avec disque de colonisation pour $m = 3$ champs.

Cette modélisation donnerait m paramètres de dispersion de graines pour chaque type d'adventice, ce qui permettrait de comparer la dispersion des différentes adventices.

Regardons maintenant les lois de probabilité du modèle graphique.

B.1 Loi de $X_{c,n+1}|GD_{c,n+1}, X_{c,n}, Y_{c,n+1}$

Dans cette section, nous allons voir plusieurs méthodes pour modéliser la loi de $X_{c,n+1}|GD_{c,n+1}, X_{c,n}, Y_{c,n+1}$.

B.1.1 Modélisation Sparse

De façon analogue au modèle sparse paramétrique, utilisons les maxima et les minima pour estimer notre probabilité de transition. On les notera :

$$m_{GD_{c,i,n+1}} = \frac{1}{\#D_{c,i}} \sum_{l=1}^{\#D_{c,i}} m_{Y_{l,n+1}}$$

et

$$M_{GD_{c,i,n+1}} = \frac{1}{\#D_{c,i}} \sum_{l=1}^{\#D_{c,i}} M_{Y_{l,n+1}}$$

A l'aide des minima et des maxima on pose

$$B_+ = \nu_0 + \nu_1 M_{Y_{c,n}} + \sum_{i=1}^{m} \exp(-\nu_3 d_{c,i}) M_{GD_{c,i,n}}$$

$$B_- = \nu_0 + \nu_1 m_{Y_{c,n}} + \sum_{i=1}^{m} \exp(-\nu_3 d_{c,i}) m_{GD_{c,i,n}}$$

où α_0, α_1 et β sont des paramètres et $\beta > 0$. La forme $\exp(-\beta d_{c,i})$ est utilisé pour obtenir une colonisation décroissante plus la distance $d_{c,i}$ est grande. La probabilité de transition est ensuite modélisée de la même manière que 4.5.1.2 64 à l'aide de B_+ et B_- définis ci-dessus. De plus le modèle peut être étendu à 4 seuils comme décrit dans la section C.

B.1.2 Modélisation Binomiale logistique

La modélisation Binomiale logistique est identique a celle décrite dans 4.5.2.1 p65. Pour les disque de colonisation on pose simplement un w_ν différent.

$$w_\nu = \nu_0 + \nu_1 X_{c,n} + \nu_2 Y_{c,n+1} + \sum_{r=1}^{3} \exp(-\nu_3 d_{c,i}) GD_{c,r,n+1}$$

Comme au-dessus, la colonisation est décroissante quand la distance $d_{c,r}$ est grande.

Annexe C

Extension du modèle Sparse

Ici nous détaillons comment augmenter le nombre d'états que la variable cachée puisse visiter. Plus précisément, pour tout $n \in \{1, \dots, N\}$ et $c \in \mathcal{C}$, on a $|X_{c,n+1}X_{c,n}| \geq 2$. Les B_+ et B_- sont définis de manière identique à précédemment.

La probabilité $\mathbb{P}(X_{c,n+1} = x_{c,n+1} | X_{c,n} = x_{c,n}, Y_{c,n+1} = y_{c,n+1}, Y^{\mathcal{C}\setminus c} = y^{\mathcal{C}\setminus c})$ avec 4 seuils est égale à

$$\begin{cases} 0 & \text{si } |x_{c,n+1} - x_{c,n}| > 2, \\ \dfrac{\max(B_+, S_{(++,X_{c,n})}) - \max(B_-, S_{(++,X_{c,n})})}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} + 2, \\ \dfrac{\min(B_+, S_{(--,X_{c,n})}) - \min(B_-, S_{(--,X_{c,n})})}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} - 2, \\ \dfrac{\min(B_+, S_{(++,X_{c,n})}) - \max(B_-, S_{(+,X_{c,n})})}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} + 1, \\ \dfrac{\min(B_+, S_{(-,X_{c,n})}) - \max(B_-, S_{(--,X_{c,n})})}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} - 1, \\ \dfrac{B_+ - S_{(-,X_{c,n})} + S_{(+,X_{c,n})} - B_- - \big(\max(B_+, S_{(+,X_{c,n})}) - \min(B_-, S_{(-,X_{c,n})})\big)}{B_+ - B_-} & \text{si } x_{c,n+1} = x_{c,n} \end{cases}$$

où les $S_{(--,X_{c,n})} < S_{(-,X_{c,n})} < S_{(+,X_{c,n})} < S_{(++,X_{c,n})}$ sont définis grâce à la loi log-normale.

Annexe D

Algorithme EM avec dépendance à la culture d'arrivé

Pour que les paramètres dépendent de la culture d'arrivée, il nous suffit d'inclure cette dépendance dans les matrices de transition, émission et les probabilités initiales. On suppose que u correspond a une culture est peut valoir $1, ..., U$. Ainsi on définit U probabilités de transition et émission différentes et $\lambda_u = (\zeta_u, \pi_u, \phi_u, A_u)$. Par conséquent il y a autant de jeux de paramètres θ, τ, μ, ν que de cultures. On définit la fonction *Crop* qui prend en argument le patch et l'année et rend la culture du patch pendant cette année. Ainsi il est possible d'appliquer l'algorithme EM au modèle avec culture. L'étape E est très similaire au modèle sans culture.

Étape E : On calcule $\tilde{\alpha}_{c,n}$ et $\tilde{\beta}_{c,n}$

$$\tilde{\alpha}_{c,n}(x_{c,n}) = \sum_{x_{c,n-1} \in \Omega_X} \tilde{\alpha}_{c,n-1}(x_{c,n-1}) \phi_{Crop(c,n)}(x_{c,n-1}, y_{c,n}, y_n^C, \lambda_{it}) A_{Crop(c,n)}(x_{c,n-1}, x_{c,n}, y_n^C, \lambda_{it})$$

avec $\tilde{\alpha}_{c,0}(x_{c,0}) = \alpha_{c,0}(x_{c,0})$ et

$$\tilde{\beta}_{c,n}(x_{c,n}) = \sum_{x_{c,n+1} \in \Omega_X} \tilde{\beta}_{c,n+1}(x_{c,n+1}) A_{Crop(c,n+1)}(x_{c,n}, x_{c,n+1}, y_{n+1}^C, \lambda_{it}) \phi_{Crop(c,n+1)}(x_{c,n}, y_{c,n+1}, y_n^C, \lambda_{it})$$

où $\tilde{\beta}_{c,N}(x_{c,N}) = \beta_{c,N}(x_{c,N})$. Alors $\rho_{c,n}(x_{c,n}) = \dfrac{\tilde{\beta}_{c,n}(x_{c,n}) \tilde{\alpha}_{c,n}(x_{c,n})}{\sum\limits_{x \in \Omega_X} \tilde{\beta}_{c,n}(x) \tilde{\alpha}_{c,n}(x)}$ et

$$\xi_{c,n}(x_{c,n-1}, x_{c,n}) = \frac{A_{Crop(c,n)}(x_{c,n-1}, x_{c,n}, y_n^C, \lambda_{it}) \tilde{\beta}_{c,n}(x_{c,n}) \tilde{\alpha}_{c,n-1}(x_{c,n-1}) \phi_{Crop(c,n)}(x_{c,n-1}, y_{c,n}, y_{n-1}^C, \lambda_{it})}{\sum\limits_{(x,x') \in \Omega_X^2} A_{Crop(c,n)}(x', x, y_n^C, \lambda_{it}) \tilde{\beta}_{c,n}(x) \tilde{\alpha}_{c,n-1}(x') \phi_{Crop(c,n)}(x', y_{c,n}, y_{n-1}^C, \lambda_{it})}$$

Étape M : L'étape M est très similaire au modèle sans culture. Voici l'espérance de la log

vraisemblance

$$
\begin{aligned}
E(X^{C,N}, Y^{C,N}, \lambda, \lambda_{it}) =& E[\ln(\mathbb{P}(Y^{C,N}, X^{C,N}|\lambda))|Y^{C,N} = y^{C,N}, \lambda_{it}] \\
=& \sum_{c=1}^{C} \ln(\zeta_{Crop(c,n)}(y_{c,0})) \\
&+ \sum_{c=1}^{C} \sum_{x \in \Omega_X} \ln(\pi_{Crop(c,0)}(x)) \rho_{c,0}(x) \\
&+ \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{(x_n, x_{n-1}) \in \Omega_X^2} \ln(A_{Crop(c,n)}(x_{n-1}, x_n, y_n^C)) \xi_{c,n}(x_{n-1}, x_n) \\
&+ \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{x_{n-1} \in \Omega_X} \ln(\phi_{Crop(c,n)}(x_{n-1}, y_{c,n}, y_{n-1}^C)) \rho_{c,n-1}(x_{n-1})
\end{aligned}
$$

L'étape de mise a jour étant identique pour chaque culture, on détaille seulement les équations des dérivées partielles de l'espérance de la log vraisemblance selon une culture u.

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \theta} = \sum_{\substack{c\in\{1,..,C\}\\ Crop(c,0)=u}} \left[y_{c,0} + (1-|\Omega_Y|)\frac{1}{1+\mathrm{e}^{-\theta}} \right]$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \tau_0} = \sum_{\substack{c\in\{1,..,C\}\\ Crop(c,0)=u}} \sum_{x_{c,0}\in\Omega_X} Q_{c,n}\rho_{c,0}(x_{c,0})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \tau_2} = \sum_{\substack{c\in\{1,..,C\}\\ Crop(c,0)=u}} \sum_{x_{c,0}\in\Omega_X} \frac{y_{c,0}}{|\Omega_Y|} Q_{c,n}\rho_{c,0}(x_{c,0})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \tau_3} = \sum_{\substack{c\in\{1,..,C\}\\ Crop(c,0)=u}} \sum_{x_{c,0}\in\Omega_X} \frac{f(y_0^{C-1})}{|f(\Omega_Y^{C-1})|} Q_{c,n}\rho_{c,0}(x_{c,0})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_0} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{(x_n,x_{n-1})\in\Omega_X^2} B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_1} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{(x_n,x_{n-1})\in\Omega_X^2} \frac{x_{c,n-1}}{|\Omega_X|} B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_2} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{(x_n,x_{n-1})\in\Omega_X^2} \frac{y_{c,n}}{|\Omega_Y|} B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \nu_3} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{(x_n,x_{n-1})\in\Omega_X^2} \frac{f(y_n^{C-1})}{|f(\Omega_Y^{C-1})|} B_{c,n}\xi_{c,n}(x_{c,n-1},x_{c,n})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_0} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{x_{c,n-1}\in\Omega_X} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_1} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{x_{c,n-1}\in\Omega_X} \frac{x_{c,n-1}}{|\Omega_X|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_2} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{x_{c,n-1}\in\Omega_X} \frac{y_{c,n-1}}{|\Omega_Y|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

$$\frac{\partial E(X^{C,N},Y^{C,N},\lambda,\lambda_{it})}{\partial \mu_3} = \sum_{\substack{(c,n)\in\{1,..,C\}\times\{1,...,N\}\\ Crop(c,n)=u}} \sum_{x_{c,n-1}\in\Omega_X} \frac{f(y_{n-1}^{C-1})}{|f(\Omega_Y^{C-1})|} B'_{c,n}\rho_{c,n-1}(x_{c,n-1})$$

où $B_{c,n} = x_{c,n} + \dfrac{1-|\Omega_X|}{1+\mathrm{e}^{w_\nu}}$, $B'_{c,n} = y_{c,n} + \dfrac{1-|\Omega_Y|}{1+\mathrm{e}^{w_\mu}}$ et $Q_{c,n} = x_{c,0} + \dfrac{1-|\Omega_X|}{1+\mathrm{e}^{z}}$.

Ensuite, il suffit de déterminer la valeur des paramètres qui annulent les dérivées partielles pour toutes les cultures.

Annexe E

Résultats supplémentaires sur les données d'Epoisses

Dans cette section nous présentons les histogrammes des flores levées simulées et des flores levées exactes des espèces *Aethusa cynapium*, *Galium aparine* et *Polygonum aviculare*.

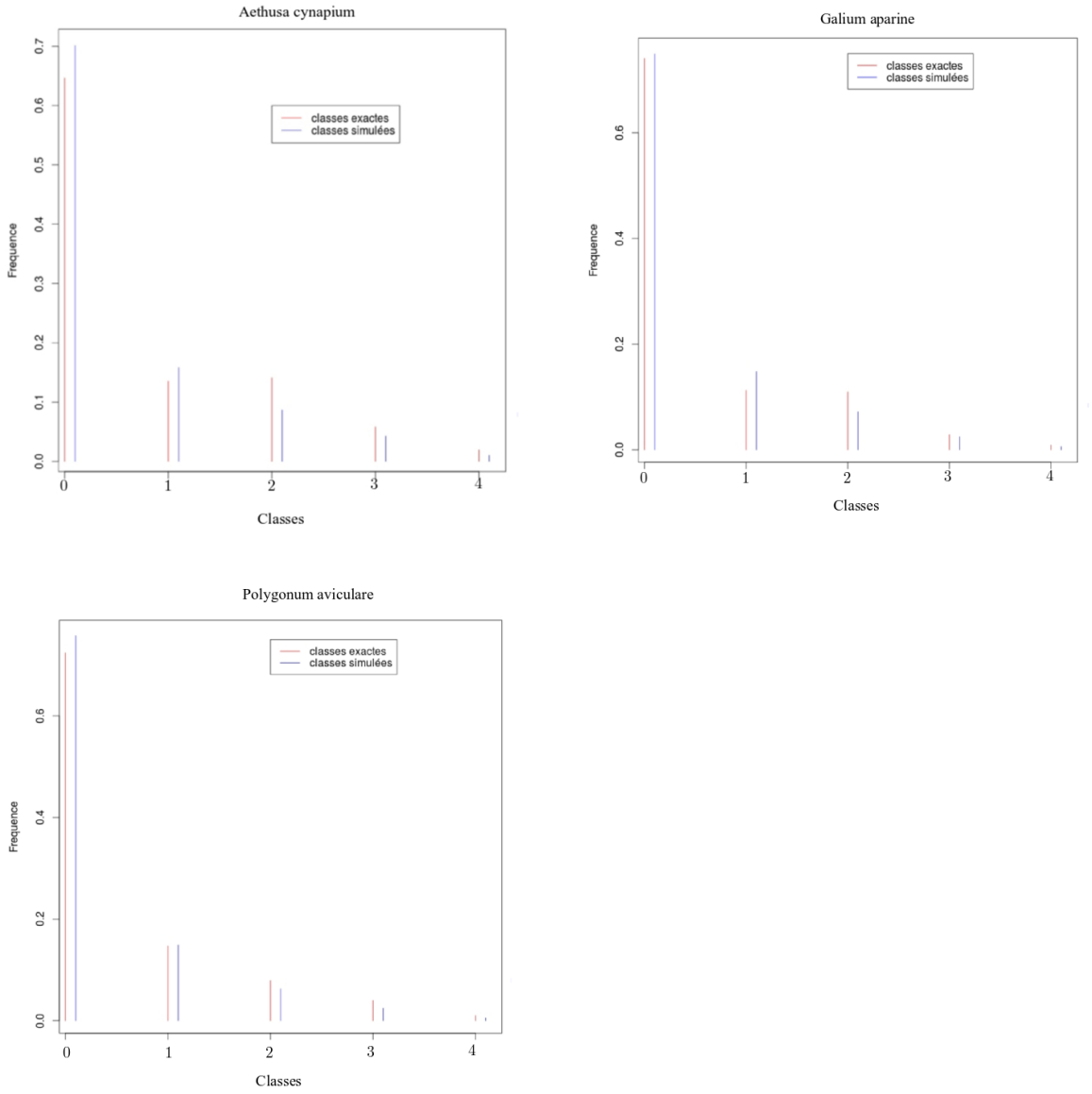


FIGURE E.1 – Histogramme des flore levée simulé et des flores exactes des espèces

Le tableau suivant montre les matrices confusion des prédictions avec le modèle selon la culture.

Alopecurus myosuroides	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	40	6	2	0	0
	1	11	7	1	0	0
	2	9	7	1	0	0
	3	1	0	3	0	0
	4	0	0	0	0	0

Aethusa cynapium	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	71	0	0	0	0
	1	1	0	0	0	0
	2	8	0	0	0	0
	3	6	0	0	0	0
	4	1	1	0	0	0

Chenopodium album	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	77	0	0	0	0
	1	0	0	0	0	0
	2	6	0	0	0	0
	3	2	0	0	0	0
	4	3	0	0	0	0

Galium aparine	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	57	4	0	0	0
	1	3	0	1	0	0
	2	14	3	1	0	0
	3	5	0	0	0	0
	4	0	0	0	0	0

Solanum nigrum	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	82	0	0	0	0
	1	2	0	0	0	0
	2	4	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0

Polygonum aviculare	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	79	0	0	0	0
	1	1	0	0	0	0
	2	5	0	0	0	0
	3	1	0	1	0	0
	4	1	0	0	0	0

Fallopia convolvulus	Etat prédit $\hat{Y}_{2017}$ / Etat exact y	0	1	2	3	4
	0	44	9	1	0	0
	1	8	3	2	0	0
	2	6	5	0	0	0
	3	6	1	2	0	0
	4	0	1	0	0	0

TABLE E.1 – Matrices de confusion des prédiction de la flore levée de l'année 2017 selon la saison des cultures

Annexe F

Les estimateurs du modèle de Pluntz dépendent de l'échelle

Dans le modèle de Pluntz et al. (2018) la survie de la banque de graines dépend de la taille de la parcelle. Ce modèle est binaire et le paramètre s correspond à la probabilité que la banque de graines survive jusqu'à l'année d'après. On suppose que toutes les graines suivent une loi de Bernouilli avec comme paramètre p correspondant à la probabilité qu'une graine survive dans la banque de graines. Ainsi $s = 1 - (1 - p)^n$ où n est le nombre de graines dans le patch. Plus le nombre de graines est grand dans la banque de graines, plus la survie de la banque de graines est élevée. Le même problème émerge avec la probabilité de germination g. Dans le modèle de Pluntz et al. (2018), g est la probabilité qu'au moins une graine germe et survive jusqu'à l'âge adulte. On suppose que q est la probabilité qu'une graine germe. De façon analogue à précédemment, on prend un champ avec $2n$ graines réparties uniformément dans tout le champ. La probabilité de germination d'au moins une graine dans tout le champ est $g = 1-(1-q)^n$ cependant la probabilité qu'une graine germe dans une moitié du champ est $1 - (1 - q)^{\frac{n}{2}}$. Ainsi l'échelle n'influence pas la valeur des paramètres si et seulement si $q = 0$.

Bibliographie

Bioflore database. `http://www2.ufz.de/biolflor/index.jsp`.

Dormancy database. `http://www.charlesgwillis.com/baskin-dormancy-database/`.

L'image de l'espèce Myrmeconema neotropicum. `https://www.berkeley.edu/news/media/releases/2008/01/16_ants.shtml`.

A. Cushman, S. and K. McGarigal (2004). Patterns in the speciesenvironment relationship depend on both scale and choice of response variables. *Oikos 105*(1), 117–124.

Adams, V. M., D. M. Marsh, and J. S. Knox (2005). Importance of the seed bank for population viability and population monitoring in a threatened wetland herb. *Biological Conservation 124*(3), 425–436.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics 16*(1), 3 – 14.

Akaike, H. (2011). *Akaike's Information Criterion*, pp. 25–25. Berlin, Heidelberg : Springer Berlin Heidelberg.

Allen, T. F. and T. W. Hoekstra (1991). Role of heterogeneity in scaling of ecological systems under analysis. In *Ecological heterogeneity*, pp. 47–68. Springer.

Allman, E. S., C. Matias, and J. A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics 37*(6A), 3099 – 3132.

Amarasekare, P. and H. Possingham (2001). Patch dynamics and metapopulation theory : the case of successional species. *Journal of Theoretical Biology 209*(3), 333 – 344.

Andersen, S. B., M. Ferrari, H. C. Evans, S. L. Elliot, J. J. Boomsma, and D. P. Hughes (2012, 05). Disease dynamics in a specialized parasite of ant societies. *PLOS ONE 7*(5), 1–8.

Arino, H. S. J., C. Aubert, L. Fontaine, J. Gall, C. Glachant, G. Johan, P. Ménétrier, C. Vacher, V. Zaganiacz, and G. Haute-Normandie (2012). *Connaître et maîtriser les adventices en grandes cultures sans herbicides*. Academic Press.

Barbu, V. S. and N. Limnios (2009). *Semi-Markov chains and hidden semi-Markov models toward applications : their use in reliability and DNA analysis*, Volume 191. Springer Science & Business Media.

BARRALIS, G., R. CHADOEUF, and J. P. LONCHAMP (1988). Longevité des semences de mauvaises herbes annuelles dans un sol cultivé. *Weed Research 28*(6), 407–418.

Baskin, C. C. and J. M. Baskin (1998). *Seeds : ecology, biogeography, and evolution of dormancy and germination.* San Diego, USA : Academic Press.

Baskin, J. M. and C. C. Baskin (2004). A classification system for seed dormancy. *Seed Science Research 14*(1), 116.

Beal., M. J. (2003). *Variational algorithm for approximate Bayesian inference*. Ph. D. thesis, M.A., M.Sci., Physics, University of Cambridge, UK,.

Behrends, E. (2000). *Introduction to Markov chains*, Volume 228. Springer.

Borgy, B., X. Reboud, N. Peyrard, R. Sabbadin, and S. Gaba (2015). Dynamics of weeds in the soil seed bank : A hidden Markov model to estimate life history traits from standing plant. *PLoS ONE 10*, e0139278.

Brand, M., N. Oliver, and A. Pentland (1997). Coupled hidden Markov models for complex action recognition. *Computer Vision and Pattern Recognition*, 994 – 999.

Bretagnolle, V. and S. Gaba (2015, Jul). Weeds for bees ? a review. *Agronomy for Sustainable Development 35*(3), 891–909.

Bulla, J. (2006). Application of hidden markov models and hidden semi-markov models to financial time series.

Bullock, J., K. Shea, and O. Skarpaas (2006). Measuring plant dispersal : an introduction to field methods and experimental design. *Plant Ecology 186*, 217–234.

Bullock, J. M., S. M. White, C. Prudhomme, C. Tansey, R. Perea, and D. A. Hooftman (2012). Modelling spread of british wind-dispersed plants under future wind speeds in a changing climate. *Journal of Ecology 100*(1), 104–115.

Byrd, R., P. Lu, J. Nocedal, and C. Zhu (1995, 9). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing 16*, 1190 – 1208.

Chadoeuf-Hannel, R. (1985). La dormance chez les semences de mauvaises herbes. *Agronomie 5*(8), 761–772.

Cohen, D. (1966). Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology 12*(1), 119 – 129.

Conn, J. S., K. L. Beattie, and A. Blanchard (2006). Seed viability and dormancy of 17 weed species after 19.7 years of burial in alaska. *Weed Science 54*(3), 464470.

Cornelisse, T. M., M. K. Bennett, and D. K. Letourneau (2013, 08). The implications of habitat management on the population viability of the endangered ohlone tiger beetle (cicindela ohlone) metapopulation. *PLOS ONE 8*(8), 1–8.

David, O., A. Garnier, C. Larédo, and J. Lecomte (2010). Estimation of plant demographic parameters from stage-structured censuses. *Biometrics 66*(3), 875–882.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Donohue, K. (2005). Seeds and seasons : interpreting germination timing in the field. *Seed Science Research 15*(3), 175187.

Dornier, A., V. Pons, and C. PO. (2011). Colonization and extinction dynamics of an annual plant metapopulation in an urban environment. *Oikos 120*(8), 1240–1246.

Dáttilo, W., J. C. F. Falcão, S. P. Yanoviak, G. Poinar, and T. J. Izzo (2013). The geographic distribution of parasite-induced fruit mimicry in cephalotes atratus (formicidae : Myrmicinae). *Journal of Parasitology 99*(1), 155–157. PMID : 22765390.

El-Keblawy, A., H. Shabana, T. Navarro, and S. Soliman (2017). Effect of maturation time on dormancy and germination of citrullus colocynthis (cucurbitaceae) seeds from the arabian hyper-arid deserts. *BMC Plant Biology 17*(1), 263.

Ferguson, J. (1980). Application of hidden markov models to text and speech. *Princeton, NJ), IDA-CRD*.

Forney, G. (1973). The Viterbi algorithm. *IEEE 63*(1), 268 – 278.

Freckleton, R. and A. Watkinson (2002). Large-scale spatial dynamics of plants : metapopulation regional ensembles and patchy populations. *Journal of Ecology 90*, 419–434.

Fréville, H., R. Choquet, R. Pradel, and P. Cheptou (2013). Inferring seed bank from hidden Markov models : new insights into metapopulation dynamics in plants. *Journal of Ecology*, 1572–1580.

Garnier, A., A. Deville, and J. Lecomte (2006). Stochastic modelling of feral plant populations with seed immigration and road verge management. *Ecological Modelling 197*(3), 373 – 382.

Gaston, K. J. (1996). Species-range-size distributions : patterns, mechanisms and implications. *Trends in Ecology & Evolution 11*(5), 197 – 201.

Ghahramani, Z. (1998). *Learning dynamic Bayesian networks*, pp. 168–197. Berlin, Heidelberg : Springer Berlin Heidelberg.

Ghahramani, Z. and M. Jordan (1997). Factorial hidden Markov models. *Mach. Learn. 29*(2-3), 245–273.

Green, J. M. (2009). Evolution of glyphosate-resistant crop technology. *Weed Science 57*(1), 108117.

Gremer, J. R. and D. L. Venable (2014). Bet hedging in desert winter annual plants : optimal germination strategies in a variable environment. *Ecology Letters 17*(3), 380–387.

Gupta, A. and B. Dhingra (2012). Stock market prediction using hidden markov models. pp. 1–4.

Gyllenberg, M. Hastings, A. and I. Hanski (1997). 5 - Structured metapopulation models. In I. Hanski and M. E. Gilpin (Eds.), *Metapopulation Biology*, pp. 93 – 122. San Diego : Academic Press.

Gyllenberg, M., A. Hastings, and I. Hanski (1997). 5 - structured metapopulation models. In *Metapopulation biology*, pp. 93–122. Elsevier.

Gómez-González, S., M. Paniw, K. Antunes, and F. Ojeda (2018). Heat shock and plant leachates regulate seed germination of the endangered carnivorous plant drosophyllum lusitanicum. *Web Ecology 18*(1), 7–13.

Han, Z., T. Lui, Q. Sun, R. Li, J. Xie, and B. Li (2014). Application of compound interest laws in biology : Reunification of existing models to develop seed bank dynamics model of annual plants. *Ecological Modelling 278*, 67–73.

Hanski, I. and C. D. Thomas (1994). Metapopulation dynamics and conservation : A spatially explicit model applied to butterflies. *Biological Conservation 68*(2), 167 – 180.

Herpigny, B. and F. Gosselin (2015). Analyzing plant cover class data quantitatively : Customized zero-inflated cumulative beta distributions show promising results. *Ecological Informatics 26*, 18 – 26.

Hurlbert, A. H. and W. Jetz (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences 104*(33), 13384–13389.

Jarry, M., M. Khaladi, M. Hossaert-McKey, and D. McKey (1995). Modelling the population dynamics of annual plants with seed bank and density dependent effects. *Acta Biotheoretica 43*(1-2), 53–65.

Kaelbling, L. P., M. L. Littman, and A. R. Cassandra (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence 101*(1), 99 – 134.

Kotliar, N. B. and J. A. Wiens (1990). Multiple scales of patchiness and patch structure : A hierarchical framework for the study of heterogeneity. *Oikos 59*(2), 253–260.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Lamy, T., O. Gimenez, J. Pointier, P. Jarne, and P. David (2013). Metapopulation dynamics of species with cryptic life stages. *the American Naturalist 181*(4), 479–491.

Levey, D. J., J. J. Tewksbury, and B. M. Bolker (2008). Modelling long-distance seed dispersal in heterogeneous landscapes. *Journal of Ecology 96*(4), 599–608.

Levin, S., D. Cohen, and A. Hastings (1984). Dispersal strategies in patchy environments. *Theoretical Population Biology 26*(2), 165–191.

Levins, R., D. Vagaggini, P. Zarattini, and G. Mura (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America 15*(3), 237–240.

MacDonald, N. and A. Watkinson (1981). Models of an annual plant population with a seedbank. *Journal of Theoretical Biology 93*(3), 643 – 653.

Manna, F., R. Pradel, R. Choquet, H. Fréville, and P. Cheptou (2017). Disentangling the role of seed bank and dispersal in plant metapopulation dynamics using patch occupancy surveys. *Ecology 98*(10), 2662–2672.

Mistro, D., L. Rodrigues, and A. Schmid (2005). A mathematical model for dispersal of an annual plant population with a seed bank. *Ecological Modelling 188*, 52–61.

Moss, S. (1985). The survival of alopecurus myosuroides huds. seeds in soil. *Weed Research 25*(3), 201–211.

Nathan, R. Muller-Landau, H. (2000). Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology & Evolution 15*(7), 278–285.

Orians, G. H. and J. F. Wittenberger (1991). Spatial and temporal scales in habitat selection. *The American Naturalist 137*, S29–S49.

Patterson, T. A., M. Basson, M. V. Bravington, and J. S. Gunn (2009). Classifying movement behaviour in relation to environmental conditions using hidden markov models. *Journal of Animal Ecology 78*(6), 1113–1123.

Patterson, T. A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos (2008). Statespace models of individual animal movement. *Trends in Ecology & Evolution 23*(2), 87 – 94.

Pluntz, M., S. Le Coz, N. Peyrard, R. Pradel, R. Choquet, and P.-O. Cheptou (2018). A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora. *Ecology Letters 0*(0).

Quintana-Ascencio, P. F., E. S. Menges, C. W. Weekley, M. I. Kelrick, and B. Pace-Aldana (2011). Biennial cycling caused by demographic delays in a fire-adapted annual plant. *Population Ecology 53*(1), 131–142.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, Volume 77, pp. 257–286.

Rabiner, L. R. and B.-H. Juang (1986). An introduction to hidden markov models. *ieee assp magazine 3*(1), 4–16.

Rai, P. and A. Lal (2016). Google PageRank Algorithm : Markov Chain Model and Hidden Markov Model . *International Journal of Computer Applications 138*(9).

Regan, T. J., M. A. McCarthy, P. W. Baxter, F. Dane Panetta, and H. P. Possingham (2006). Optimal eradication : when to stop looking for an invasive plant. *Ecology letters 9*(7), 759–766.

Renaud, V., R. François, K. Yutaka, O. Isabelle, and G. Sylvain (2013). The joint evolution of dispersal and dormancy in a metapopulation with local extinctions and kin competition. *Evolution 67*(6), 1676–1691.

Rimey, R. D. and C. M. Brown (1991, Nov). Controlling eye movements with hidden markov models. *International Journal of Computer Vision 7*(1), 47–65.

Rydén, T. et al. (2008). Em versus markov chain monte carlo for estimation of hidden markov models : A computational perspective. *Bayesian Analysis 3*(4), 659–688.

Rydén, T. (2008). Em versus markov chain monte carlo for estimation of hidden markov models : a computational perspective. *3*(4), 659–688.

Saas, Y. and F. Gosselin (2014). Comparison of regression methods for spatially-autocorrelated count data on regularly-and irregularly-spaced locations. *Ecography 37*(5), 476–489.

Schneider, O., J. Roger-Estrade, J.-N. Aubertot, and T. Doré (2006). Effect of seeders and tillage equipment on vertical distribution of oilseed rape stubble. *Soil and Tillage Research 85*(1), 115 – 122.

Seglias, A.and Williams, E., A. Bilge, and A. Kramer (2018). Phylogeny and source climate impact seed dormancy and germination of restoration- relevant forb species. *Plos one 13*(2), e019193.

Simpson, G. (1990). Seed dormancy in grasses. *Cambridge University Press*.

Stokstad, E. (2013). The war against weeds down under. *Science 341*(6147), 734–736.

Taab, A. (2009). *Seed Dormancy and Germination in Solanum nigrum and S. physalifolium as Influenced by Temperature Conditions*. Ph. D. thesis, Swedish University of Agricultural Sciences Uppsala.

Taylor, K. (1999). Galium aparine l. *Journal of ecology 87*(4), 713–730.

Turner, M. G., R. V. O'Neill, R. H. Gardner, and B. T. Milne (1989). Effects of changing spatial scale on the analysis of landscape pattern. *Landscape ecology 3*(3-4), 153–162.

Ueno, K. (2002). Effects of desiccation and a change in temperature on germination of immature grains of wheat (triticum aestivum l.). *Euphytica 126*(1), 107 – 113.

Usher, M. (1981). Modelling ecological succession, with particular reference to markovian models. In *Vegetation dynamics in grasslans, healthlands and mediterranean ligneous formations*, pp. 11–18. Springer.

Venable D, L. (2007). Bet hedging in a guild of desert annuals. *Ecology 88*(5), 1086–1090.

Venable D, L. and J. S. Brown (1988). The selective interactions of dispersal, dormancy, and seed size as adaptations for reducing risk in variable environments. *The American Naturalist 131*(3), 360–384.

Vleeshouwers, L. M., H. J. Bouwmeester, and C. M. Karssen (1995). Redefining seed dormancy : An attempt to integrate physiology and ecology. *Journal of Ecology 83*(6), 1031–1037.

Wainwright, M. and M. Jordan (2008). Graphical models, exponential families, and variational inference. In *Foundations and Trends in Machine Learning*, Volume 1, pp. 1–305.

Wilson, J. S. and A. D. Worsham (1988). Combinations of nonselective herbicides for difficult to control weeds in no-till corn, zea mays, and soybeans, glycine max. *Weed Science 36*(5), 648–652.

Yanoviak, S. P., M. Kaspari, R. Dudley, and G. Poinar Jr (2008). Parasite-induced fruit mimicry in a tropical canopy ant. *The American Naturalist 171*(4), 536–544.

Yedidia, J. S., W. T. Freeman, and Y. Weiss (2003). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium 8*, 236–239.

Yu, S.-Z. (2010). Hidden semi-markov models. *Artificial intelligence 174*(2), 215–243.

Zhang, R., J. Baskin, C. Baskin, Q. Mo, L. Chen, X. Hu, and Y. Wang (2017). Effect of population, collection year, after-ripening and incubation condition on seed germination of stipa bungeana. *Scientific reports 7*(1), 13893.