Analyse probabiliste, étude combinatoire et estimation paramétrique pour une classe de modèles de croissance de plantes avec organogenèse stochastique

Cédric Loi

ÉCOLE CENTRALE DES ARTS ET MANUFACTURES
« ÉCOLE CENTRALE PARIS »

Thèse
présentée par Cédric LOI
pour l'obtention du GRADE DE DOCTEUR

Spécialité : Mathématiques Appliquées
Laboratoire d'accueil : Mathématiques Appliquées aux Systèmes (MAS)

Analyse probabiliste, étude combinatoire et estimation paramétrique pour une classe de modèles de croissance de plantes avec développement stochastique

Soutenue le 31/05/2011
Devant un jury composé de :

Mme. Mireille RÉGNIER	(Rapporteur)
Mme. Florence D'ALCHÉ-BUC	(Rapporteur)
M. Jacques LÉVY-VÉHEL	(Président)
Mme. Amélie MATHIEU	(Examinatrice)
M. Paul-Henry COURNÈDE	(Directeur)

N° ordre : 2011ECAP0023

• **Résumé** :

Dans cette thèse, nous nous intéressons à une classe particulière de modèles stochastiques de croissance de plantes structure-fonction à laquelle appartient le modèle GreenLab. L'objectif est double. En premier lieu, il s'agit d'étudier les processus stochastiques sous-jacents à l'organogenèse. Un nouveau cadre de travail combinatoire reposant sur l'utilisation de grammaires formelles a été établi dans le but d'étudier la distribution des nombres d'organes ou plus généralement des motifs dans la structure des plantes. Ce travail a abouti à la mise en place d'une méthode symbolique permettant le calcul de distributions associées à l'occurrence de mots dans des textes générés aléatoirement par des L-systèmes stochastiques. La deuxième partie de la thèse se concentre sur l'estimation des paramètres liés au processus de création de biomasse par photosynthèse et de son allocation. Le modèle de plante est alors écrit sous la forme d'un modèle de Markov caché et des méthodes d'inférence bayésienne sont utilisées pour résoudre le problème.

Mots-clés : modèle de croissance de plantes structure-fonction, GreenLab, processus de branchement, méthode symbolique, modèle de Markov Caché, inférence bayésienne, filtrage particulaire

• **Abstract** :

This PhD focuses on a particular class of stochastic models of functional-structural plant growth to which the GreenLab model belongs. First, the stochastic processes underlying the organogenesis phenomenon were studied. A new combinatorial framework based on formal grammars was built to study the distributions of the number of organs or more generally patterns in plant structures. This work led to the creation of a symbolic method which allows the computation of the distributions associated to word occurrences in random texts generated by stochastic L-systems. The second part of the PhD tackles the estimation of the parameters of the functional submodel (linked to the creation of biomass by photosynthesis and its allocation). For this purpose, the plant model was described by a hidden Markov model and Bayesian inference methods were used to solve the problem.

Keywords : functional-structural plant growth model, GreenLab, branching process, symbolic method, hidden Markov model, Bayesian inference

Remerciements :

Dans ces quelques lignes, je tiens à remercier l'ensemble des personnes qui ont contribué à cette thèse et/ou qui m'ont accompagné durant toutes ces années.

Je remercie tout d'abord Jacques Lévy-Véhel, Mireille Régnier, Florence d'Alché-Buc et Amélie Mathieu d'avoir accepté de participer à ma soutenance et de l'intérêt qu'ils ont porté à ma thèse.

Je souhaite remercier Jean Françon qui est à l'origine des travaux effectués en combinatoire et dont les conseils m'ont été d'une grande aide. Je voudrais également rendre hommage à Philippe Flajolet qui a été une véritable source d'inspiration pour ces travaux de thèse.

Je remercie l'école doctorale de l'ECP sous la direction de Jean-Hubert Schmitt et le laboratoire MAS dirigé successivement par Christian Saguez, Etienne De Rocquigny et Frédéric Abergel pour les moyens qui ont été mis à ma disposition.

Merci également aux secrétaires INRIA Christine Biard, Isabelle Biercewicz et Delphine Goyer pour la gestion des aspects administratifs et en particulier un grand merci à Sylvie Dervin, secrétaire du laboratoire MAS, pour son aide et sa bonne humeur au quotidien.

Un grand merci à Paul-Henry pour son encadrement durant toutes ces années (depuis l'étude en autonomie de fin de première année à Centrale jusqu'à la fin de la thèse !). C'est grâce à lui que j'ai développé mon goût pour la recherche et l'application des mathématiques à la biologie. Je tiens particulièrement à le remercier pour sa patience et son sens de l'écoute qui m'ont permis de trouver ma voie professionnelle.

Ce fut un grand plaisir de vivre ces années de thèse dans l'équipe Digiplante. J'y ai rencontré des personnes fort sympathiques qui m'ont aidé dans mon travail de recherche : Philippe De Reffye pour son initiation au monde de la botanique et de la modélisation, Véronique Letort pour le partage de ses connaissances en modélisation de la croissance des plantes et la relecture du deuxième chapitre de la thèse, Samis Trevezas qui m'a beaucoup appris en statistiques et qui a relu les chapitres six et sept et enfin Vincent Le Chevalier pour nos différents débats concernant mes travaux.

Je remercie également les amis doctorants et doctorantes que j'ai connus durant ces quatre années au laboratoire et avec lesquels je partage d'excellents souvenirs que ce soit les moments passés au bureau, les pauses café, les soirées jeux/pizza et autres : Vincent, Véronique, Aurélie, Mahendra, Qi Rui, Charlotte, Takuya, Marc, Sylvain ainsi que tous les autres doctorants de l'équipe Digiplante et associés. Merci aussi à tous les amis "hors Centrale" qui m'ont accompagné pendant ma thèse : Sébastien, Tristan, Le Seg, Guitou et Nicolas entre autre.

Un grand merci à ma famille et en particulier à mes parents, mon frère, ma grand-mère et Kiki qui m'ont soutenu tout au long de ma thèse et qui m'ont permis d'en arriver là où j'en suis aujourd'hui. Merci également à toute la famille Gauthray-Guyénet pour leur soutien.

Enfin et surtout un énorme merci à Lucile pour m'avoir accompagné étape par étape durant cette thèse. Elle a su me réconforter dans les moments de doute et me donner la force de continuer jusqu'au bout. Je lui dois beaucoup.

Table des matières

Table des figures

Liste des tableaux

Liste des algorithmes

Chapitre 1

Contexte et problématiques

L'homme ne saurait subsister sans les plantes. Elles jouent un rôle fondamental au quotidien : production de nourriture que ce soit par les cultures en champs ou sous serre, régulateur environnemental de par le cycle naturel du carbone, production de biocarburant avec le colza ou la betterave ... Face à la situation économique et environnementale actuelle, il apparaît essentiel de gérer au mieux leur exploitation. En effet, la population mondiale ne cesse de croître tandis la quasi-totalité des terres arables est déjà exploitée. Il est donc capital d'augmenter la rentabilité des surfaces agricoles tout en limitant l'empreinte écologique. Ainsi, la compréhension des mécanismes de croissance des plantes et celle de leurs interactions avec l'environnement est un des enjeux majeurs du $21^{\text{ème}}$ siècle.

C'est dans un tel contexte que s'est développée la modélisation de la croissance des plantes. L'objectif principal de ces modèles est de pouvoir reproduire le comportement des plantes dans un environnement donné. Parmis les nombreuses applications, nous pouvons citer la prédiction qualitative et quantitative de la production végétale ou encore l'optimisation des cultures par rapport à des ressources énergétiques limitées. La fin des années 1990 a connu l'essor d'une nouvelle classe de modèle de croissance de plante : les modèles structure-fonction (voir Sievänen et al. (2000), Prusinkiewicz (2004), Godin and Sinoquet (2005) et De Reffye et al. (2008)). Ces modèles ont pour but de simuler le développement de la structure de la plante ainsi que son fonctionnement biophysique (c'est-à-dire la production de biomasse[1] par photosynthèse et son allocation). Afin de reproduire la croissance le plus fidèlement possible, ils intègrent dans leur formalisme des processus biophysiques de plus en plus complexes. Les mathématiques deviennent donc un outil indispensable pour analyser et exploiter ces modèles.

Quoi de plus intéressant pour un chercheur que d'explorer de nouvelles contrées regorgeant de problématiques complexes, variées et passionantes pouvant aboutir sur des applications très prometteuses. C'est dans cet état d'esprit que j'ai effectué ma thèse dans l'équipe Digiplante au laboratoire MAS[2] de l'École Centrale Paris (ECP). Cette thèse fut l'occasion de découvrir le domaine de la botanique mais également d'approfondir (ou apprendre) de nombreuses disciplines des mathématiques. Les chapitres suivants

[1] la biomasse est la matière végétale composant les plantes, voir le chapitre 2

[2] Mathématiques Appliquées aux Systèmes

traitent ainsi de questions de nature diverse que ce soit en modélisation, en probabilité, en combinatoire ou encore en statistique.

Dans ce chapitre introductif, nous présentons d'abord l'équipe Digiplante puis le contexte scientifique de la thèse. Ensuite, nous donnons les différents axes de développement du manuscrit.

1.1 L'équipe Digiplante

1.1.1 Présentation

Digiplante[3] est une équipe projet INRIA cofondée en 2002 par Philippe de Reffye (CIRAD[4]) et Paul-Henry Cournède (ECP[5]). Ses axes de recherche sont centrés sur la modélisation mathématiques de la croissance des plantes et ses applications en agronomie et en foresterie. Plus précisément, l'équipe développe un modèle de croissance de plante intitulé GreenLab (voir le paragraphe suivant) dont les objectifs sont les suivants :

- Prévision quantitative et qualitative de la production végétale des cultures et des forêts ;
- Contrôle Optimal des cultures et des modes d'exploitation ;
- Optimisation du processus d'amélioration génétique.

1.1.2 Le modèle GreenLab

GreenLab est un modèle de croissance de plante structure-fonction (voir le chapitre 2 pour une description complète du modèle). Il intègre donc dans son formalisme le développement de la structure de la plante et son fonctionnement caractérisé par les processus de création de biomasse par photosynthèse et son allocation aux organes de la plante. Les équations de GreenLab ont été publiées pour la première fois en 2003 (voir de Reffye and Hu (2003)). Par la suite, plusieurs versions ont vu le jour (voir le chapitre 2 pour plus de détails). Le modèle GreenLab 2 (Kang et al. (2004)) se caractérise par un modèle de développement stochastique. Dans le modèle GreenLab 3 (Mathieu (2006)), il y a une rétroaction du fonctionnement sur le développement. Le modèle GreenLab 4 (voir par exemple Pallas et al. (2009)) couple les caractéristiques des deux modèles GreenLab 2 et 3.

Le modèle GreenLab est générique, c'est-à-dire qu'il peut simuler la croissance de toute sorte de plante que ce soit des petites plantes ou des grands arbres. Dans cette optique, les processus impliqués doivent être suffisamment généraux pour être présents chez toutes les plantes mais représentés le plus simplement possible (tout en conservant un sens botanique). Le modèle est donc décrit de façon concise par un faible nombre d'équations ce qui le rend exploitable facilement du point de vue mathématique (estimation des paramètres, contrôle optimal, analyse de sensibilité ...). La paramétrisation du

[3] http ://digiplante.saclay.inria.fr/

[4] centre de Coopération Internationale en Recherche Agronomique pour le Développement. http ://www.cirad.fr/

[5] École Centrale Paris. http ://www.mas.ecp.fr/website/

modèle s'en trouve également simplifiée ce qui fait sa force par rapport à la majorité des autres modèles de croissance structure-fonction. En effet, dans GreenLab, l'estimation des paramètres se fait à l'échelle de la plante entière et non processus par processus. Les paramètres estimés intègrent ainsi les nombreuses interactions complexes des différents processus biophysiques de la croissance des plantes. On peut parler de modèle intégratif.

Le modèle a été adapté sur toute sorte de plantes qu'elles soient agronomiques (le maïs, Guo et al. (2006) ; la betterave, Lemaire (2010)), tropicales (*Cecropia sciadophylla*, Letort et al. (2009)) ou même des arbres (le hêtre, Letort et al. (2008a) ; le pin, Guo et al. (2007)).

1.1.3 Les logiciels

Quatre principaux logiciels basés sur le modèle de croissance GreenLab ont été implémentés à l'heure actuelle :

- GreenScilab[6] : il s'agit d'un logiciel open-source développé par le LIAMA. Plus précisément, il se présente sous la forme d'une *toolbox* Scilab. Le logiciel ne prend en compte que les versions GreenLab 1 et 2 du modèle. De par sa licence libre, il est un excellent outil de diffusion du modèle GreenLab et est également utilisé dans des sessions de travaux pratiques sur la modélisation de la croissance des plantes.

- Digiplante (logiciel) : logiciel éponyme de l'équipe, il est implémenté en C++ afin d'optimiser les temps de calcul pour la simulation. Les versions GreenLab 1 et 3 du modèle sont codées dans le logiciel. De par ses fonctionnalités, il est plus performant que le logiciel GreenScilab. Il offre en effet des représentations 3D plus réalistes et est bien adapté pour les aspects calibration du modèle. Digiplante est le logiciel le plus utilisé par l'équipe et un grand nombre de partenaires pour la plupart des applications (CIRAD-AMAP[7], ITB[8], CAU[9], CAF[10], INRA Grignon ...).

- Pygmalion : il s'agit d'une plateforme développée très récemment permettant à son utilisateur de créer son propre modèle de croissance. De nombreux outils mathématiques sont également mis à disposition pour exploiter ce modèle : identification paramétrique, contrôle optimal, analyse de sensibilité ...

- DGP SDK : ce logiciel se concentre principalement sur la simulation et la visualisation 3D des plantes que ce soit à l'échelle individuelle ou à celle du paysage.

[6] http ://www.scilab.org/products/modules/pem/greenscilab

[7] UMR botAnique et bioinforMatique de l'Architecture des Plantes. http ://amap.cirad.fr/fr/index.php

[8] Institut Technique de la Betterave. http ://www.itbfr.org/

[9] China Agricultural University. http ://www.cau.edu.cn/cie/en/

[10] Chinese Academy of Forestry. http ://www.caf.ac.cn/

1.2 Contexte scientifique

Les travaux de cette thèse ont été motivés par les résultats de recherches sur la structure des plantes et les processus biophysiques permettant sa mise en place. L'étude botanique de la structure des plantes est récente. Les premiers travaux remontent aux années 1970 avec entre autres ceux de Hallé and Oldeman (1970) et Hallé et al. (1978). Parallèlement au développement de l'informatique, de nombreuses recherches ont ensuite été menées pour l'obtention de structures 3D réalistes. C'est dans ce contexte qu'ont été créés des modèles de plante dits architecturaux (voir Kurth (1994a)). Plusieurs procédés ont été employés pour construire ces représentations 3D. Parmis ceux-ci, l'utilisation de L-systèmes reste l'un des plus populaires. Les L-systèmes sont des grammaires à réécriture parallèle opérant sur des chaînes de caractères[11]. A l'origine, ils ont été introduits en 1968 par Lindenmayer pour modéliser le développement d'organismes multicellulaires simples, cf Lindenmayer (1968). Ce n'est que dans les années 1980 qu'ils ont été envisagés pour générer des structures de plantes, cf Aono and Kunii (1984) et Smith (1984). Par la suite, les L-systèmes ont connu un développement important dans ce domaine (voir Prusinkiewicz and Lindenmayer (1990)). Afin de proposer des structures de plus en plus réalistes, plusieurs déclinaisons ont alors vu le jour telles que les L-systèmes stochastiques, les L-systèmes « context-sensitive » et les « relational growth grammars » (pour ces derniers voir Kurth (1994b) et Kurth (2007)). Les L-systèmes ont également été utilisés dans le but de simuler le développement de la structure des plantes pour certains modèles de croissance structure-fonction (voir par exemple le modèle ADEL pour le maïs, Fournier and Andrieu (1999), le coton avec Hanan (2004) et le modèle L-PEACH pour les arbres, Allen et al. (2005)).

L'étude mathématique des structures de plante a été plus tardive. de Reffye (1979) a été parmi les premiers à modéliser les mécanismes biophysiques de l'organogenèse[12] du caféier avec des processus aléatoires. Les différents modules composant la structure d'une plante sont alors mis en place suivant un ensemble de règles probabilistes. La plante toute entière devient donc un objet aléatoire. Des calculs de distributions, de moyennes et de variances ont également été effectués notamment pour des organes occupant une position particulière dans la configuration de la plante.

De nombreuses études probabilistes sur la structure des plantes ont ensuite été menées par le groupe AMAP depuis les années 1990. de Reffye et al. (1988), Costes et al. (1992) proposent des modèles d'organogenèse utilisant la théorie du renouvellement. Pour ce type de modélisation, il est possible de déterminer la distribution du temps d'apparition entre deux phytomères[13] successifs le long d'un axe, cf Costes et al. (1993). Des recherches ont ensuite été conduites sur la façon dont les phytomères se succèdent le long d'un axe donné d'une plante, chacun de ces phytomères étant caractérisé par un ensemble de données botaniques. Par exemple, Costes et al. (1994) s'intéressent à la répartition des phytomères le long d'un axe en fonction de la présence ou non de fruit sur les phytomères en questions. Pour résoudre le problème, les axes ont été imaginés

[11] voir le chapitre 4 pour une définition et une étude complète des L-systèmes

[12] processus de création d'organe

[13] La structure des plantes est un assemblage d'entités élémentaires appelées phytomères, voir chapitre 2

comme étant la trajectoire d'une chaîne de Markov (l'état initial de la chaîne est donné par le phytomère à la base de l'axe). Chaque phytomère est alors caractérisé par un état qui est la présence de fruit ou non. D'autres données peuvent être prises en compte pour ce genre d'approche comme le nombre d'axes latéraux portés par un phytomère, voir Heuret et al. (2002). L'étude des caractéristiques botaniques des phytomères le long d'un axe a surtout été traitée par Yann Guédon. Son objectif est de repérer les zones homogènes d'un axe, c'est-à-dire des portions d'axe pour lesquelles les phytomères qui les composent présentent des caractéristiques botaniques similaires. Pour ce faire, il utilise la théorie des chaînes de semi-Markov cachées (Kulkarni (1995)). Chaque axe est considéré comme la trajectoire d'une chaîne de semi-Markov cachée. Les phytomères sont caractérisés par un état ontogénique caché qui est observable par l'intermédiaire de ses caractéristiques botaniques (voir Guédon et al. (2003)). Le temps de séjour des phytomères dans un état ontogénique donné est représenté par une autre variable aléatoire qui est également caché *a priori*. Ces travaux ont non seulement été appliqués sur de vraies plantes (Heuret et al. (2003) et Guédon (2005)) mais ont abouti également à des résultats mathématiques portant sur l'estimation des chaînes de semi-Markov cachées (Guédon (2003)) et l'analyse statistique de ces chaînes (par exemple quel est l'état le plus probable, cf Guédon (2007)). Ces recherches ont ensuite été étendues aux arbres de Markov cachés pour lesquels c'est la structure de la plante toute entière qui est utilisée et non uniquement les axes (voir Durand et al. (2005)). Cette approche permet alors de tenir compte de phénomènes globaux à l'échelle de la plante.

L'étude de l'agencement des phytomères le long des axes a permis d'améliorer les modèles d'organogenèse de certains modèles de croissance de plante structure-fonction. C'est le cas du modèle L-PEACH qui utilise les L-systèmes pour générer ses structures de plante. Un nouveau modèle L-PEACH a ainsi vu le jour avec un modèle d'organogenèse couplant L-systèmes et chaînes de semi-Markov cachés (voir Lopez et al. (2008)). Le modèle obtenu est beaucoup plus flexible et les résultats de simulation concordent mieux avec les plantes observées.

Inspirée par les travaux de de Reffye (1979) et de Jaeger and de Reffye (1992), une version stochastique du modèle structure-fonction GreenLab (dénommée GreenLab 2) a été créée au début des années 2000 (voir Kang et al. (2003) et Kang et al. (2008)). A chaque étape de croissance, le comportement des bourgeons est dicté par un ensemble de processus aléatoires (processus de survie, de dormance, nombre aléatoire d'axes latéraux portés par un phytomère, ...). L'espérance et la variance du nombre de phytomères pour une plante d'un âge donné ont été calculées explicitement grâce à la théorie des processus composés (Kang et al. (2008)). De plus, les équations du modèle ont permis le calcul de l'espérance et la variance de la quantité de biomasse produite à un instant donné. Kang et al. (2007) ont également écrit les fonctions génératrices associées au nombre de phytomères dans une plante d'un âge donné grâce à une approche combinatoire. Ainsi, la plante est codée par un mot donnant sa composition. L'organogenèse est alors caractérisée par un ensemble de règles de production donnent les évolutions de chacune des lettres composant le mot. L'idée d'utiliser la combinatoire dans le cadre d'études mathématiques pour la croissance des plantes n'est pas une idée nouvelle. Françon (1990) imaginait déjà tout le potentiel de la représentation de la structure d'une plante par un arbre binaire ce qui permettrait d'utiliser notamment des résultats et méthodes d'analyse

combinatoire (voir Flajolet and Sedgewick (2009)). La vision d'une plante en tant que structure combinatoire a également été perçue par Viennot et al. (1989) pour des besoins d'ordre graphique.

1.3 Problématiques et axes de développement

L'étude des structures des plantes et de leur mise en place est récente et de nombreux problèmes n'ont pas encore été résolus. Le point qui ressort le plus est la multitude des approches concernant la représentation de l'organogenèse. En effet, un certain nombre de formalismes ont été introduits dans le but de simuler et d'étudier le développement de la structure parmi lesquels nous pouvons citer la représentation par des automates (cf Zhao et al. (2001)), les processus de branchement multitypes (voir Kang et al. (2007) et Loi and Cournède (2008)) ou encore les L-systèmes (cf Prusinkiewicz and Lindenmayer (1990)). Ces différentes approches présentent pourtant de très fortes similitudes. Il serait donc intéressant de les comparer, d'établir des liens et de caractériser leurs différences pour montrer ce que chacune d'entre elles peut apporter.

Un autre point est celui de la nature des objets étudiés. Les études probabilistes se sont concentrées soit sur des calculs de distributions du nombre d'organes d'un type donné soit sur l'agencement des phytomères suivant un ensemble de caractéristiques botaniques. La question de l'occurrence de motifs particuliers dans la structure d'une plante n'a pas encore été abordée. Le terme motif désigne par exemple un agencement topologique précis comme un embranchement[14].

Il est également clair que tout le potentiel de la combinatoire dans le cadre des plantes n'a pas encore été dévoilé. Le lien entre plante et arbre planaire est intuitif. Il y a donc beaucoup de résultats à obtenir à partir du vaste champs mathématique qu'est la combinatoire.

Enfin, dans le cadre du modèle de croissance GreenLab 2, il n'existe pas de méthode d'estimation générale pour les paramètres liés au fonctionnement de la plante.

Dans le présent manuscrit, nous tentons de répondre aux problèmes évoqués ci-dessus. La thèse porte sur l'étude d'une classe de modèles de croissance de plante avec développement stochastique. Cette classe est représentée par un métamodèle se caractérisant par deux processus biophysiques : le développement de la plante (*i.e.* la mise en place de sa structure) et le fonctionnement (*i.e.* la création de biomasse par photosynthèse et son allocation aux organes de la plante). Le métamodèle est alors étudié d'abord sous un angle probabiliste puis sous un angle combinatoire. Par le biais de ces travaux, nous verrons les similitudes des différentes approches de modélisation existante et ce que chacune d'entre elles peut apporter. Ces résultats sont également utilisés pour l'estimation des paramètres du métamodèle que ce soit au niveau de l'organogenèse ou du fonctionnement.

Le métamodèle est introduit au chapitre 2. L'objectif est d'écrire un modèle suffisament général pour représenter une classe de modèles de croissance de plante. En premier lieu, nous présentons le développement de la plante comme le couplage de

[14] la notion de motif est bien illustrée dans le chapitre 5

deux phénomènes biologiques : l'organogenèse et la différenciation. Les processus de photosynthèse et d'allocation liés au fonctionnement sont ensuite détaillés. Le modèle de croissance GreenLab est ensuite présenté comme un cas particulier du métamodèle et sera utilisé comme exemple d'application dans la suite de la thèse.

Dans le chapitre 3, nous présentons et étudions un modèle de développement stochastique pour le métamodèle. L'organogenèse est alors le résultat d'un ensemble de processus aléatoires dictant le devenir des bourgeons à chaque étape de croissance. Les phase-types multivariées sont utilisées pour modéliser les changements d'état (*i.e.* la différenciation) des bourgeons apicaux des axes. Une étude probabiliste du modèle d'organogenèse est ensuite conduite. La dynamique d'évolution des bourgeons est identifiée comme un processus de branchement multitype et les fonctions génératrices associées au nombre d'organe de tout type sont écrites. La finitude de la croissance de la plante est également étudiée.

Le monde de la croissance végétale est mis de côté pour celui de la combinatoire dans le chapitre 4. Nous y présentons en effet une méthode qui trouvera son utilité dans le chapitre suivant. L'objectif est le calcul de la distribution associée au nombre d'occurrences d'un mot donné dans un texte généré aléatoirement par un L-système stochastique. Pour ce faire, une méthode symbolique dans la tradition Flajolet (voir Flajolet and Sedgewick (2009)) est développée pour l'étude des ensembles de mots pondérés. Un structure de semi-anneau est établie afin de pouvoir écrire une spécification appropriée. Nous donnons également deux théorèmes de décomposition qui contribuent à l'obtention de cette spécification. La méthode symbolique est ensuite illustrée au travers de deux exemples.

Un nouveau cadre de travail combinatoire est établi pour le modèle de développement stochastique dans le chapitre 5. La structure d'une plante est codée par un mot de Dyck et les règles de production de l'organogenèse sont données par un L-système stochastique. Grâce aux résultats du chapitre 4, il devient alors possible de déterminer la distribution associée au nombre d'occurrences d'un motif précis dans la structure de la plante. Nous déterminons par exemple la distribution du nombre d'apex ou encore de structures en Y. Le chapitre se termine par la mise en place d'une méthode d'estimation permettant de retrouver les probabilités intervenant dans le modèle de développement stochastiques à partir de données botaniques liées à la structure de la plante.

Les chapitres 6 et 7 sont dédiés à l'estimation des paramètres liés au fonctionnement du métamodèle. Ils apparaissent comme une application des résultats obtenus dans les trois précédents chapitres. Le chapitre 6 est purement bibliographique et porte sur un ensemble de méthodes d'inférence bayésiennes pour des modèles de Markov cachés. Nous présentons d'abord les modèles statistiques mis en jeu et nous montrons comment les adapter pour pouvoir faire l'estimation des paramètres qui interviennent dans les modèles. Quatre méthodes d'inférence sont ensuite présentées : le filtre de Kalman sans parfum, le filtre particulaire, le filtre particulaire convolé et le filtre particulaire Rao-Blackwellisé. Dans le chapitre 7, nous appliquons ces méthodes pour l'estimation des paramètres liés au fonctionnement du métamodèle. Ce dernier est alors présenté sous la forme d'un modèle de Markov caché. Nous donnons également un ensemble de recommandations permettant la mise en œuvre pratique des méthodes. Celles-ci sont ensuite analysées et comparées au travers de trois cas-tests reposant sur le modèle GreenLab. Nous proposons également des estimateurs pour les bruits de modélisation et de mesure

du modèle ce qui, par bootstrap paramétrique, nous permet d'avoir la distribution *a posteriori* des paramètres et les intervalles de confiance associés. Le chapitre se termine par une application sur données réelles avec le cas de la betterave sucrière.

Remarques préliminaires à la lecture du manuscrit :

- Les différents termes botaniques employés dans la thèse sont résumés dans un glossaire, voir page 197 ;
- Tous les symboles introduits dans le manuscrit sont répertoriés à la page 199 ;
- Les chapitres 3, 4 et 5 (partie probabiliste et combinatoire) sont indépendants des chapitres 6 et 7 (partie statistique). La lecture du chapitre 2 est indispensable pour la compréhension des deux parties.

Chapitre 2

Description d'une classe de modèles de croissance de plantes

Les travaux effectués dans cette thèse sont valables pour tout une classe de modèles de croissance de plantes. L'objectif de cette section est de présenter cette classe par le biais d'un métamodèle noté $\mathcal{M}$. Celui-ci intègre dans son formalisme deux phénomènes biophysiques intervenant dans la plupart des modèles structure-fonction (voir Le Roux et al. (2001)) :

- le développement : mise en place de la structure de la plante ;
- le fonctionnement : ensemble des processus de création de biomasse[1] par photosynthèse et son allocation aux organes de la plante.

Ces processus sont décrits au niveau macroscopique (c'est-à-dire à l'échelle de l'organe). Nous commençons ce chapitre par la présentation du développement de la plante (section 2.1). Le fonctionnement est ensuite détaillé dans la section 2.2. Un certain nombre de modèles structure-fonction correspond à cette description dont le modèle de croissance GreenLab introduit dans la section 2.3. Ce modèle sevira d'exemple d'application pour la thèse.

N.B. 2.1 Le modèle de développement de $\mathcal{M}$ sera complété par un modèle de développement stochastique $\mathcal{S}$ introduit dans le chapitre 3.

2.1 Développement

Le développement d'une plante désigne le processus de mise en place de sa structure. Dans cette thèse, le développement est le résultat de deux phénomènes biologiques :

- l'organogenèse : création de nouveaux organes par la plante ;
- la différenciation : changement de classe physiologique du méristème apical d'un axe (voir la section 2.1.4).

[1] matière végétale composant la plante

Les trois premières sections sont dédiées à l'organogenèse (sections 2.1.1, 2.1.2 et 2.1.3). Plus précisément, nous donnons quelques principes de modélisation permettant de la simplifier et de l'étudier mathématiquement. La section 2.1.4 explique plus en détails la différenciation. Ensuite, le concept de structure est défini plus formellement, cf section 2.1.5. La section 2.1.6 introduit des variables liées à la composition de la plante. Celles-ci vont jouer un rôle important pour la suite de la thèse. La dernière section s'intéresse à la représentation des règles de production du développement sous forme d'automates.

2.1.1 Méristème et bourgeon

L'acteur principal de l'organogenèse est le méristème primaire. Il s'agit d'un tissu végétal constitué de cellules indifférenciées. Ces cellules vont se diviser au cours de la mitose et se différencier par la suite pour former de nouveaux organes (feuilles, fleurs, fruits, ...). Ce sont donc les méristèmes primaires qui construisent la structure de la plante. On distingue principalement deux types de méristèmes primaires suivant leur position sur la plante :

- Méristème apical (également appelé apex) : ce méristème est situé au bout d'un axe (une tige, une branche ou une racine par exemple). Il est alors responsable de l'expansion en longueur de cet axe.
- Méristème axillaire : celui-ci est situé à l'aisselle des feuilles. Dans ce cas, il peut être à l'origine d'un nouvel axe (= axe latéral).

La figure 2.1 positionne les méristèmes primaires à l'échelle de la plante.

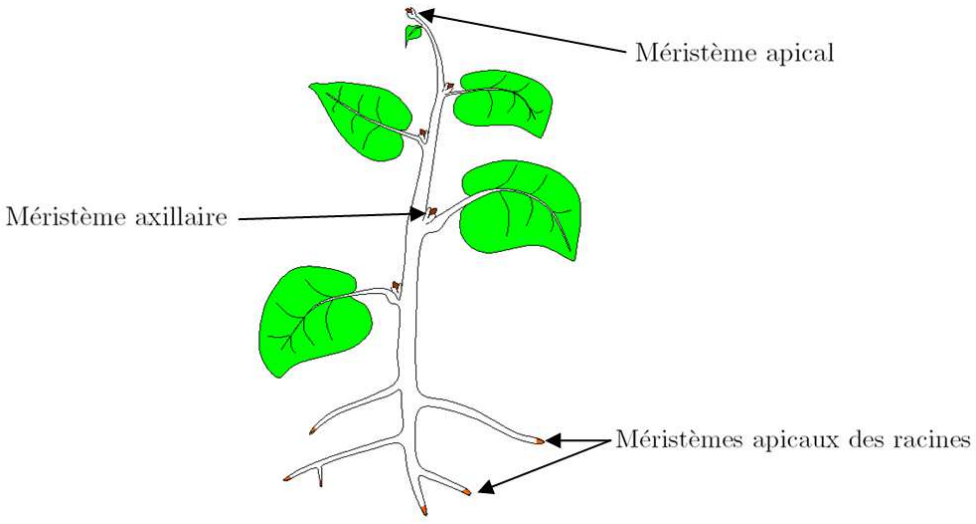

FIG. 2.1 – Localisation des méristèmes primaires d'une plante.

Suivant le mode de fonctionnement du méristème, la croissance de la plante peut être continue ou rythmique. Il y a croissance continue lorsque le méristème fonctionne en permanence. La plante crée donc de nouveaux organes en continue. C'est le cas de

nombreuses herbacées et de certains arbres tropicaux (le genre *Cecropia* par exemple, Letort et al. (2009)). Pour d'autres plantes, le méristème marque des pauses dans son fonctionnement. Ces pauses sont généralement synchronisées avec des fluctuations environnementales mais restent parfois observées également si la plante est placée en environnement constant (Barthélémy and Caraglio (2007)). Dans ce cas, la croissance de la plante est dite rythmique. Par exemple, la plupart des arbres en zone tempérée sont à croissance rythmique. A l'approche de l'hiver, le méristème cesse de se développer et reprend son activité au printemps. Ainsi, le cycle de développement dure un an (voir la figure 2.4).

En ce qui concerne les plantes à croissance rythmique, les méristèmes primaires se situent à l'intérieur des bourgeons. Pendant la période hivernale, des organes (les primordia) se forment dans les bourgeons. Au printemps, ces derniers éclosent et les organes apparaissent. Il est également possible que des organes non préformés dans le bourgeon apparaîssent par la suite de manière continue au cours de l'année. On parle alors de néoformation. On observe également, comme chez le hêtre (Nicolini (1997)), que la croissance de la pousse annuelle peut être marquée d'une ou plusieurs interruptions pendant lesquelles peuvent se former des bourgeons fugaces. On parle alors de mono- ou polycyclisme. Dans cette thèse, nous considérons seulement l'échelle de la pousse annuelle.

Par la suite, afin de simplifier le processus d'organogenèse, nous ne ferons pas de distinction entre le bourgeon et le méristème qu'il contient. Les organes responsables de l'organogenèse seront alors désignés de façon générale par le terme bourgeon. Le vocabulaire utilisé pour les méristèmes leur sera également appliqué (bourgeon apical, bourgeon axillaire, ...). Les bourgeons n'existent en réalité que chez les arbres. Chez les plantes vertes par exemple, le méristème n'est pas contenu dans un bourgeon. Du point de vue modélisation, nous supposerons cependant que ceux-ci sont contenus dans des bourgeons « fictifs » (voir Mathieu (2006)). L'ensemble sera alors également désigné par le terme bourgeon.

2.1.2 Discrétisation spatiale de la structure d'une plante

Phytomère

La plante peut être considérée comme un assemblage de structures élémentaires appelées phytomères (voir White (1979) et Barthélémy et al. (1997) ou encore la figure 2.5 pour un exemple). La façon dont les phytomères sont reliés les uns aux autres constitue la topologie de la plante. Un phytomère est un ensemble d'organes composé essentiellement :

- d'un entrenœud (partie de la tige située entre deux nœuds successifs, c'est-à-dire deux points d'insertion de feuilles successifs) ;
- d'une ou plusieurs feuilles constituées d'un limbe (surface de la feuille captant la lumière) et d'un pétiole (tige raccordant le limbe à l'entrenœud) ;

et potentiellement :

– d'un bourgeon apical (aussi appelé bourgeon terminal) ;
– de bourgeons axillaires (également appelés bourgeons latéraux) ;
– de fleurs ;
– de fruits.

D'autres organes peuvent être rattachés au phytomère. Par exemple, les racines sont considérées comme un organe attaché au phytomère situé à la base de la plante. La figure 2.2 donne un schéma d'un phytomère.

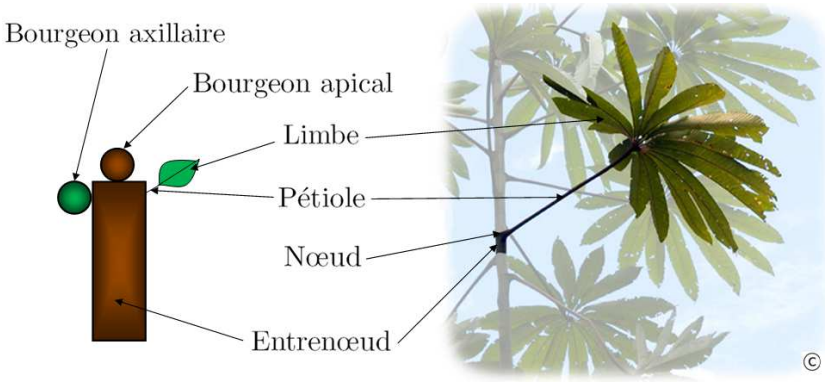

FIG. 2.2 – Description d'un phytomère. Le phytomère de droite provient d'un *Cecropia*, Letort et al. (2009) (le dessin est de Patrick Heuret).

Classes physiologiques

Le méristème apical d'un axe peut produire des phytomères ayant des caractéristiques morphologiques (c'est-à-dire liées à son apparence externe) ou fonctionnelles différentes en fonction de ce que l'on appelle son stade de différenciation ou âge physiologique (voir Barthélémy and Caraglio (2007)). Il est alors possible de regrouper les phytomères émis par les méristèmes de même âge physiologique en classes, que nous appellerons classes physiologiques (ou CP en abrégé). Chaque classe physiologique est alors caractérisée par l'âge physiologique qui lui correspond. Pour la suite de la thèse, nous désignerons par classe physiologique d'âge k l'ensemble des organes associés à l'âge physiologique k.

La classe physiologique 1 est attribuée à l'axe situé à la base de la plante (le tronc pour un arbre par exemple). Le stade final de différenciation se dénote par $CP_m \in \mathbb{N}^*$ et est attribué en général à des axes courts, fins et ne portant pas de branche. Pour la plupart des plantes, un nombre maximal de 5 ou 6 CPs est suffisant pour classer tous les axes (voir par exemple les catégories d'axes définies pour le cèdre *Cedrus Atlantica*, Sabatier and Barthélémy (1999)). Sauf quelques exceptions rares qui ne seront pas étudiées dans la thèse, un axe de classe physiologique i ne peut porter que des axes latéraux de classe physiologique j supérieure (c'est-à-dire $i \leq j$, voir par exemple la figure 2.3). Lorsque $i = j$, on parle de réitération.

N.B. 2.2 La graine sera considérée comme un bourgeon de classe physiologique 1.

N.B. 2.3 Attention à ne pas confondre classe physiologique et ordre de ramification. L'ordre de ramification se définit comme suit : l'ordre 1 est attribué à l'axe principal (le tronc par exemple) ; lorsqu'un axe latéral est porté par un axe d'ordre k, celui-ci est d'ordre $k+1$. Il existe certaines plantes chez lesquelles ordre de ramification et classes physiologiques coïncident (le pin par exemple, Guo et al. (2007), ou encore la figure 2.6) et d'autres non (le hêtre, Letort et al. (2008b), et la figure 2.3).

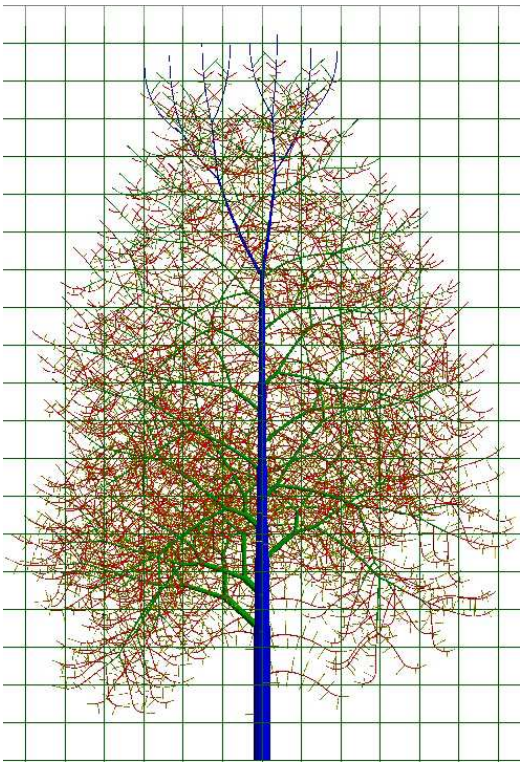

FIG. 2.3 – Illustration des différentes classes physiologiques d'une plante. La classe 1 est en bleu, la 2 en vert et la 3 en rouge. Dans cet exemple, classes physiologiques et ordres de ramification ne coïncident pas.

2.1.3 Discrétisation temporelle de l'organogenèse

Cycle de développement

Dans la plupart des modèles de croissance de plante, le processus d'organogenèse est discrétisé en temps (voir par exemple les modèles LIGNUM Perttunen et al. (1996), GreenLab , Yan et al. (2004), et ADEL, Fournier and Andrieu (1998)). La croissance de la plante peut alors s'écrire sous la forme d'un système dynamique discret. Par la suite, nous appellerons cycle de développement le pas de temps associé. Dans cette thèse, nous le définissons de différentes façons suivant le type de croissance considéré. Dans le cas d'une plante à croissance rythmique, le cycle de développement est défini comme le temps séparant l'apparition de deux pousses successives (par exemple, le cycle de développement est d'un an pour les arbres en zone tempérée). Si la croissance est

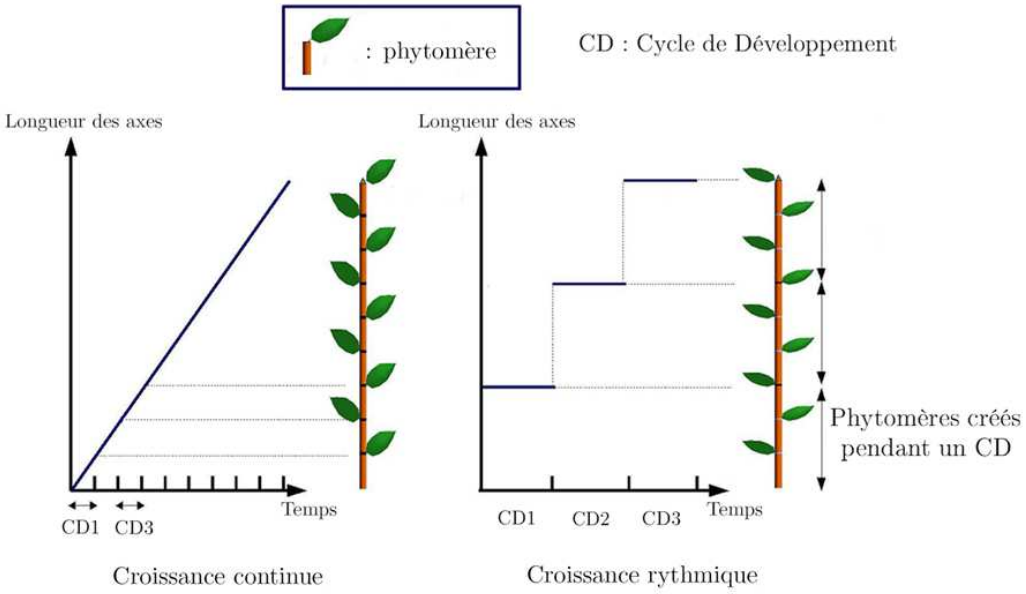

FIG. 2.4 – Croissance rythmique/continue et cycle de développement. Dans l'exemple ci-dessus, un bourgeon produit trois phytomères par cycle de développement dans le cas rythmique.

continue, le cycle de développement est simplement défini comme le temps nécessaire au méristème apical de l'axe principal pour former un nouveau phytomère (voir la figure 2.4). Dans ce cas, le cycle de développement est plus court et peut varier de quelques jours à quelques semaines. Par convention, le cycle de développement zéro marque le début de la croissance (la plante est alors à l'état de graine).

La discrétisation temporelle de l'organogenèse permet ainsi de simplifier considérablement sa modélisation. Que la croissance soit continue ou rythmique, le cycle de développement sera décrit de la façon suivante : au début du cycle, les bourgeons de la plante éclosent pour donner naissance à de nouveaux phytomères. Ces phytomères possèdent des bourgeons apicaux et potentiellement des bourgeons axillaires. Ces bourgeons pourront alors éclore au début du cycle de développement suivant et engendrer de nouveaux phytomères.

Par la suite, nous appellerons unité de croissance l'ensemble des phytomères produit par un bourgeon durant un cycle de développement.

Notion d'âge

L'âge d'un axe ou d'un organe est donné par le nombre de cycles de développement écoulés depuis sa création. Par convention, l'âge zéro est attribué aux axes ou organes qui viennent d'être créés (les bourgeons entre autres ; dans ce cas, la graine sera considérée comme un bourgeon de classe physiologique 1 et d'âge 0). Plus généralement, l'âge d'une plante correspond au nombre de cycles de croissance écoulés depuis la germination de la graine.

2.1.4 Différenciation

En général, un bourgeon apical est de la même classe que le phytomère qui le porte. Cependant, il est possible que ce bourgeon change de classe (il se différencie) sous l'effet du vieillissement. Ce phénomène est connu sous le nom de différenciation (ou encore de mutation). Un bourgeon apical de CP i ne peut se différencier que vers une classe j supérieure ($j > i$). Le stade ultime de la différenciation correspond en général à la mort du bourgeon ou à sa transformation en fleur.

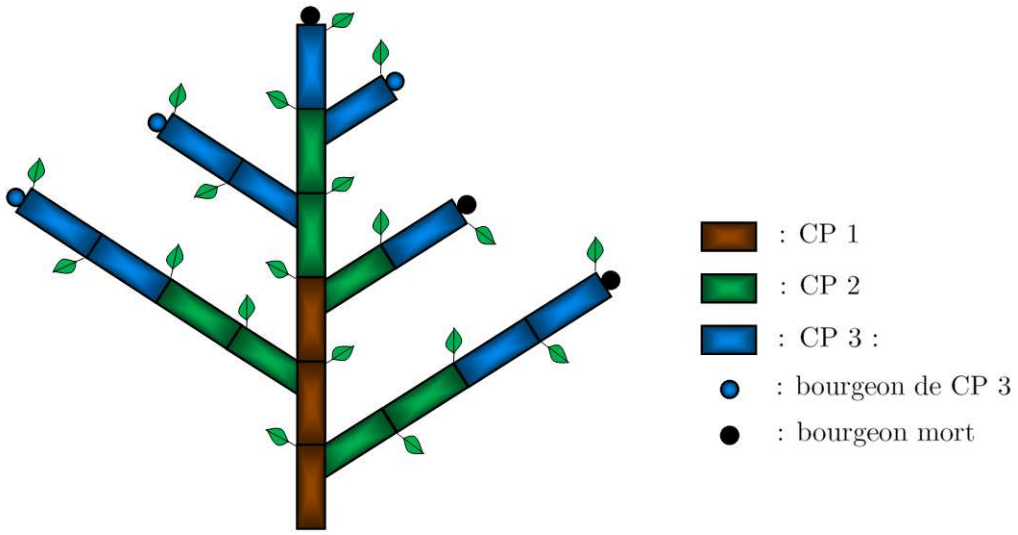

FIG. 2.5 – Exemple de différenciation. Sur l'axe principal, le bourgeon apical passe de la classe physiologique 1 à 2 après 3 phytomères, de la classe 2 à 3 après 2 autres et de la classe 3 à un bourgeon mort après un dernier.

2.1.5 Structures

Une structure de classe physiologique i et d'âge k est la structure formée par l'ensemble des phytomères issus d'un bourgeon de CP i après k cycles de développement. Autrement dit, celle-ci peut être vue comme une « sous-plante » d'âge k dont la graine serait un bourgeon de CP i. En particulier, une structure de classe physiologique i et d'âge 0 est un bourgeon de CP i. Une plante d'âge k est une structure de CP 1 et d'âge k. La figure 2.6 montre quelques exemples de structures.

Le concept de structure simplifie considérablement l'étude théorique et la simulation de la croissance. En effet, il est possible de décomposer la structure d'une plante en un ensemble de structures plus jeunes. De façon générale, une structure de CP i et d'âge k peut se décomposer en une unité de croissance contenant des phytomères de CP i et un ensemble de structures de CP $j \geq i$ et d'âge $k - 1$. Cette description récursive de la topologie débouche sur de nombreuses applications. Entre autre, elle est la clé de voûte de nombreuses démonstrations concernant l'étude de la structure d'une plante et de son évolution. Elle permet aussi un gain au niveau du temps de calcul pour la simulation de la croissance d'une plante, la complexité des algorithmes passant d'exponentielle à polynomiale (voir la méthode des sous-structures, Cournède et al. (2006)).

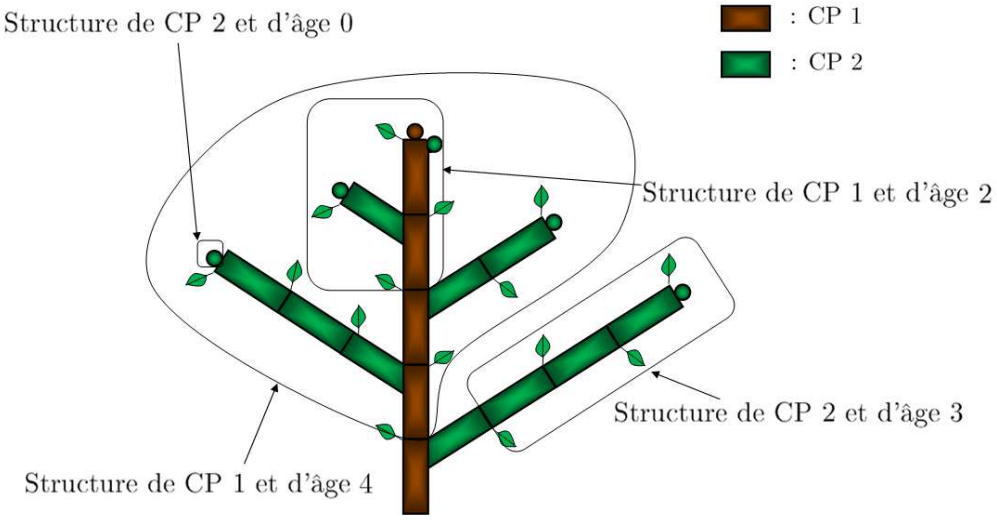

FIG. 2.6 – Exemples de structures. La plante ci-dessus est une structure de CP 1 et d'âge 5.

2.1.6 Variables topologiques

Les variables topologiques sont liées à la composition de la plante à un cycle de développement donné. Afin de les introduire, nous associons à chaque sorte d'organe un symbole. Ainsi, les lettres b, e, l, p, f_r, f_l et r sont associées respectivement aux bourgeons, entrenœuds, limbes, petioles, fruits, fleurs et racines. Parfois, il n'est pas nécessaire de distinguer les parties pétiole et limbe des feuilles. Dans ce cas, les symboles l et p sont remplacés par un unique symbole f qui désigne la feuille toute entière. Si un organe est également caractérisé par une classe physiologique, celle-ci vient alors compléter le symbole le décrivant. Par exemple, un entrenœud de CP i sera représenté par le symbole e_i. Par la suite, nous noterons $\mathcal{B}$ l'ensemble des symboles décrivant les bourgeons d'une plante donnée :

$$\mathcal{B} = \{b_i\}_{i\in\{1,\dots,CP_m\}}$$

et $\mathcal{O}$ l'ensemble des symboles permettant de décrire les autres organes ($\mathcal{O}$ ne contient pas nécessairement tous les autres symboles possibles mais seulement ceux associés aux organes pouvant potentiellement faire partie de la plante. Par exemple, si une plante ne donne pas de fruit, le symbole f_r ne sera pas présent dans $\mathcal{O}$).

N.B. 2.4 $\mathcal{B}$ représente en fait l'ensemble des organes permettant l'extension de la structure de la plante et qui sont amenés à évoluer au fil des cycles de développement tandis que $\mathcal{O}$ représente, lui, les organes constituant son architecture et qui ne changent pas.

Pour tout $k \in \mathbb{N}$ et $a \in \mathcal{B}\cup\mathcal{O}$, soit N_k^a la variable donnant le nombre total d'organes a au début du cycle de développement k (c'est-à-dire avant l'étape d'organogenèse du même cycle). Ainsi, si $a = b \in \mathcal{B}$, N_k^b donne le nombre de bourgeons b vivants au début du cycle k et qui seront en activité lors de l'organogenèse du même cycle (il s'agit des bourgeons susceptibles de produire de nouveaux phytomères au cycle k).

N.B. 2.5 Pour tout organe $o \in \mathcal{O}$, nous supposons par convention que la variable N_k^o comptabilise tous les organes de type o créés depuis le cycle de développement 0, y compris ceux qui sont morts.

N.B. 2.6 Les organes $o \in \mathcal{O}$ créés au cours de l'organogenèse du cycle de développement k sont comptabilisés dans N_{k+1}^o et non dans N_k^o. Le nombre d'organes $o \in \mathcal{O}$ créés lors de l'organogenèse du cycle de développement k vaut $N_{k+1}^o - N_k^o$. Ces organes sont issus de bourgeons qui sont actifs lors de l'organogenèse du même cycle (bourgeons comptabilisés dans la variable N_k^b).

N.B. 2.7 Etant donné que la graine est considérée comme un bourgeon de CP 1 et d'âge 0 au cycle de développement 0, on a $N_0^b = \delta_{b_1}(b)$ avec δ_{b_1} le symbole de Kronecker centré en b_1. Les organes $o \in \mathcal{O}$ issus de la germination de la graine sont comptabilisés dans N_1^o.

Pour $k \in \mathbb{N}$, soit N_k le vecteur de $\mathbb{N}^{card(\mathcal{B})+card(\mathcal{O})}$ donnant la composition de la plante au début du cycle k (avant l'étape d'organogenèse) :

$$N_k = (N_k^a)_{a \in \mathcal{B} \cup \mathcal{O}}, \qquad k \geq 0, \tag{2.1}$$

Le processus $(N_n)_{n \geq 0}$ joue un rôle important dans l'organogenèse et sera étudié dans le chapitre 3. Enfin, pour $(n_1, n_2) \in \mathbb{N}^2$ avec $n_1 < n_2$, soit $N_{n_1 \to n_2}$ le vecteur de $\mathbb{N}^{(n_2 - n_1 + 1)card(N_0)}$ donnant l'évolution de la composition de la plante du cycle n_1 au cycle n_2 :

$$N_{n_1 \to n_2} = (N_{n_1}, N_{n_1+1}, \ldots, N_{n_2}) \tag{2.2}$$

$N_{0 \to n}$ représente l'évolution de la structure de la plante d'âge n depuis sa création et intervient dans les processus écophysiologiques de la croissance (voir la section 2.2 sur le fonctionnement).

2.1.7 Représentation par des automates

Au début de chaque cycle de développement, les bourgeons actifs de la plante peuvent éclore pour donner de nouveaux phytomères. Nous appellerons règles de production l'ensemble des évolutions possibles associées à chaque type de bourgeon. Si le développement est déterministe, alors un bourgeon d'un type donné n'a qu'une seule évolution possible. En revanche, si il est stochastique, ce même bourgeon peut avoir plusieurs évolutions possibles. Dans ce cas, une probabilité d'occurrence est associée à chacune de ces évolutions. Le développement stochastique de la plante sera présenté et étudié dans le chapitre 3.

Les règles de production définissent un ensemble d'axiomes permettant de construire la structure topologique de la plante cycle de développement après cycle de développement. Dans toute la thèse, nous faisons l'hypothèse que les bourgeons se comportent indépendamment les uns des autres. Concernant l'organogenèse, il existe plusieurs manières de représenter ces règles. La représention la plus classique (et la plus visuelle) est celle sous forme d'automates à états finis (Prusinkiewicz and Lindenmayer (1990), Zhao et al. (2003)). Ces automates donnent le devenir de chaque bourgeon de la plante

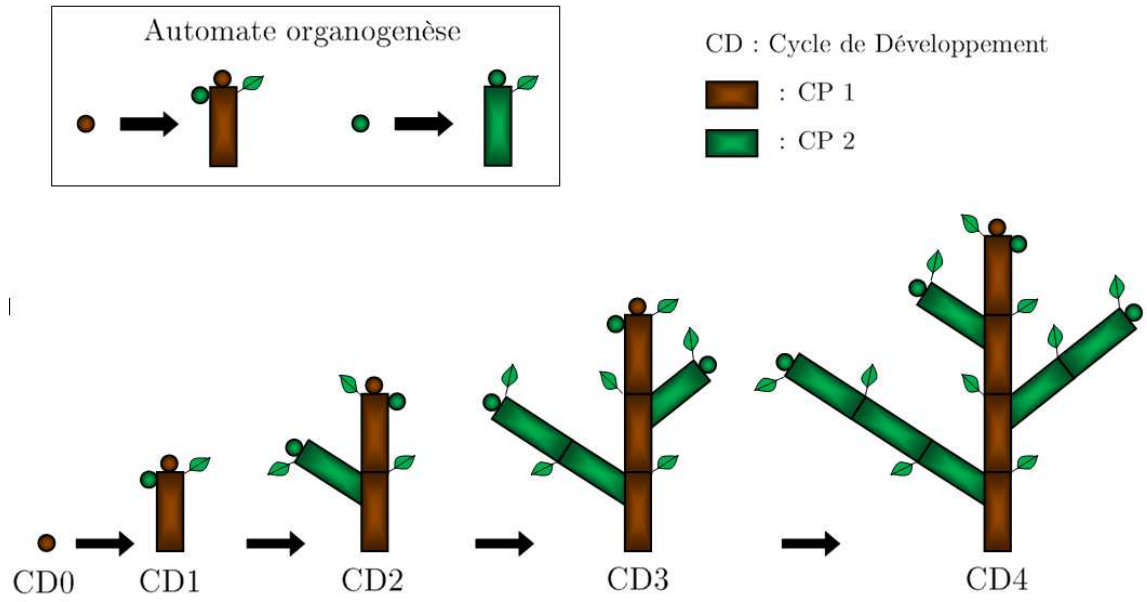

FIG. 2.7 – Exemple de règles de production déterministes. Notons que, contrairement à ce qui est indiqué par les automates, les axes latéraux alternent quant à leur position par rapport à l'axe principal. Cette représentation est plus naturelle et ne change rien quant à la suite de la thèse.

après une étape d'organogenèse. Supposons que nous disposons d'une plante d'âge n. La plante d'âge $n+1$ s'obtient alors en remplaçant chaque bourgeon de la plante d'âge n par l'une de ses évolutions données par les règles de production des automates. La figure 2.7 donne des exemples d'automates déterministes.

N.B. 2.8 Le terme automate désigne ici un ensemble de règles permettant de construire automatiquement la structure de la plante. Il ne correspond pas exactement à la définition standard d'un automate utilisée en informatique théorique (voir Flajolet and Sedgewick (2009)).

N.B. 2.9 Afin d'alléger l'écriture des règles de production, seuls seront représentés les automates des organes pouvant évoluer d'un cycle de développement à l'autre. Par convention, les organes dont les automates d'évolution ne sont pas représentés sont considérés inchangés (du point de vue de leur nature botanique) au cours de la croissance de la plante (les entrenœuds par exemple).

Il est également possible de représenter les règles de différenciation (appelée aussi mutation, voir la section 2.1.4) des bourgeons apicaux sous forme d'automate. Le chiffre au-dessus de chaque flèche indique le nombre de cycles après lequel a lieu le changement de classe physiologique. A chaque cycle de développement, les bourgeons de la plante évoluent donc non seulement selon les règles données par les automates de l'organogenèse mais aussi selon celles données par les automates associés au phénomène de différenciation. Il est donc nécessaire d'instaurer une priorité entre ces deux processus. Par convention, la priorité est donnée à l'organogenèse. En conséquence, à un cycle de

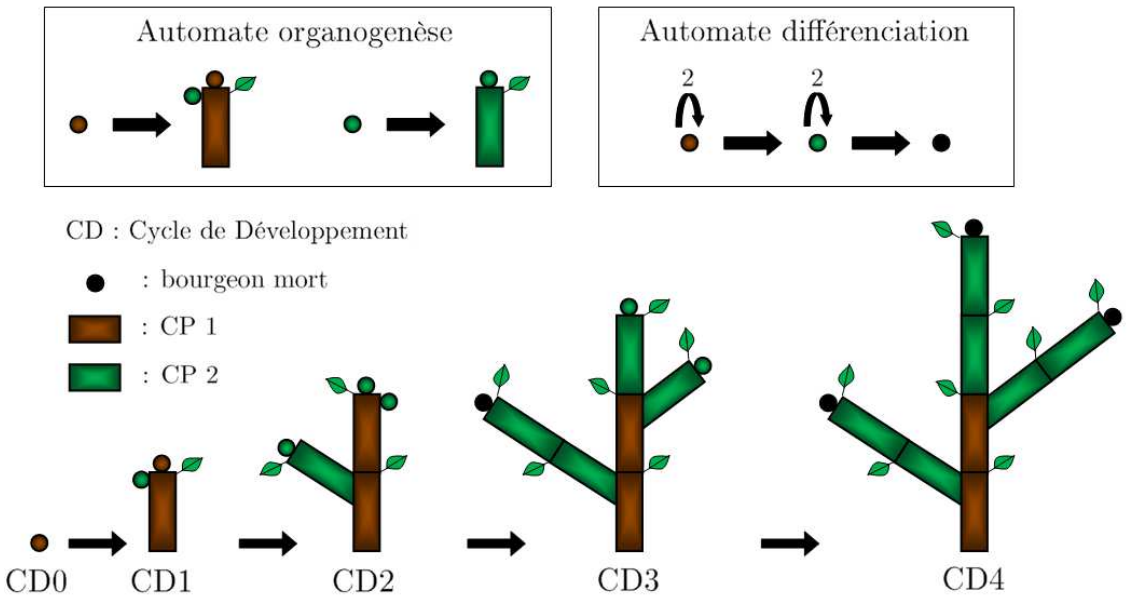

FIG. 2.8 – Exemple de règles de production couplées avec de la différenciation. La priorité est accordée aux règles de l'organogenèse.

développement donné, les bourgeons vont d'abord évoluer suivant les règles de production de l'organogenèse et ensuite suivre les règles de différenciation (voir la figure 2.8 pour un exemple).

Il existe d'autre façons de représenter les règles de production. La représentation sous forme de grammaire formelle est très présente dans le milieu botanique (voir Smith (1984), Viennot et al. (1989), Françon (1990) ou Kurth (1996)). La structure de la plante est donnée par une succession de lettres appelée mot. La grammaire formelle donne alors un ensemble de règles (établies à partir des règles de production) permettant de faire évoluer le mot (et donc la structure de la plante). Une grammaire formelle fréquemment utilisée est le L-système, Prusinkiewicz and Lindenmayer (1990) (voir le chapitre 5 pour une description complète).

2.2 Fonctionnement

Le fonctionnement de la plante est le résultat d'un certain nombre de processus écophysiologiques. L'écophysiologie est le domaine de la biologie qui étudie le comportement des organismes en interaction avec leur environnement. En ce qui concerne les plantes, un phénomène écophysiologique de première importance est la photosynthèse. Durant la photosynthèse, l'eau et les sels minéraux du sol ainsi que le dioxide de carbone présent dans l'air sont transformés en assimilats (sucres) grâce à l'énergie fournie par la lumière. Au niveau des feuilles, la sève brute se transforme en sève élaborée. Celle-ci contient les assimilats créés qui vont alors être distribués à l'ensemble des organes en expansion (= allocation). Ces assimilats seront désignés par la suite sous le terme de biomasse. Le fonctionnement de la plante représente ainsi l'ensemble des processus de

création de biomasse par photosynthèse et son allocation aux organes en expansion.

L'objectif de cette section est de présenter le modèle retenu pour le fonctionnement du métamodèle $\mathcal{M}$. Les phénomènes impliqués sont en réalité très complexes mais nous allons adopter ici un point de vue macroscopique qui nous permet de présenter un cadre très générique pour leur modélisation. Les hypothèses de validité du modèle de fonctionnement sont les suivantes :

- le fonctionnement est discrétisé en temps et le pas de temps associé est celui du développement (*i.e.* le cycle de développement) ;
- la production de biomasse dépend de la surface foliaire, que ce soit à l'échelle de la plante ou à l'échelle individuelle ;
- l'allocation se fait à l'échelle de l'organe ;
- la production de biomasse et son allocation ne dépendent pas de la topologie de la plante.

Le modèle de croissance TOMSIM dédié à la tomate (voir Heuvelink (1996) et Heuvelink (1999)) peut s'inscrire dans un tel cadre. Le modèle GreenLab également, voir la section 2.3.

La section 2.2.1 présente le modèle associé à la production de biomasse et la section 2.2.2 celui associé à son allocation. Ensuite, nous expliquons comment ces phénomènes s'intègrent dans le cycle de développement (cf section 2.2.3). La plupart des fonctions présentées par la suite dépendent d'un vecteur de paramètres Θ_{fonc} caractérisant le fonctionnement d'une plante donnée. Θ_{fonc} est inconnu *a priori* et doit être estimé à partir de données expérimentales (voir le chapitre 7).

2.2.1 Photosynthèse

Nous présentons le modèle associé à la production de biomasse par photosynthèse. Nous rappelons que, afin de simplifier le métamodèle $\mathcal{M}$ et l'étude qui en sera faite par la suite, le processus de photosynthèse est discrétisé en temps et que le pas de temps choisi est le cycle de développement (le même pas de temps que celui de l'organogenèse). Pour $n \in \mathbb{N}^*$, soit Q_n la quantité totale de biomasse créée au cours du cycle de développement n. Q_n est donnée par l'équation de production suivante (aussi appelée équation de photosynthèse) :

$$Q_n = \Phi_n(S_n^{act}, E_n, \Theta_{fonc}), \qquad n \geq 1. \tag{2.3}$$

avec Φ_n une fonction borélienne. S_n^{act} est un vecteur dont les composantes donnent les surfaces de toutes les feuilles de la plante au cours du cycle n. Le vecteur E_n traduit l'impact de l'environnement sur la quantité de biomasse créée. Suivant les modèles de croissance de plante considérés, E_n incorpore les effets liés à la température de l'air, l'ensoleillement et l'apport en eau. E_n est donc le résultat d'un ensemble de données expérimentales associées à l'environnement de la plante.

N.B. 2.10 Au cycle de développement 0, la plante est à l'état de graine. La quantité de biomasse Q_0 disponible au cours du même cycle n'est alors pas produite par photosynthèse mais est contenue dans la graine.

2.2.2 Allocation de la biomasse

La biomasse créée Q_n est ensuite distribuée à l'ensemble des organes en expansion au cycle de développement n. Pour tout $o \in \mathcal{O}$, soit T^o_{exp} le temps d'expansion associé à un organe de type o. Celui-ci correspond au nombre de cycles de développement durant lequels un organe o consomme de la biomasse. Ainsi, un organe o créé au cycle n (et comptabilisé dans N^o_{n+1}) reçoit durant son expansion de la biomasse en provenance de Q_n pour la première fois et de $Q_{n+T^o_{exp}-1}$ pour la dernière fois.

La quantité de biomasse reçue par un organe dépend de son type $o \in \mathcal{O}$, de son âge k et du cycle de développement n en cours. Elle dépend aussi de $N_{0\to n+1}$ car l'allocation dépend du nombre total d'organes en expansion durant le cycle de développement actuel (la biomasse doit être partagée entre les différents organes en expansion). Pour tout $o \in \mathcal{O}$ et $l \in \{0, \ldots, T^o_{exp} - 1\}$, soit $Al^{o,l}_n(Q_n, N_{0\to n+1}, \Theta_{fonc})$ la quantité de biomasse allouée à un organe o d'âge l au cycle de développement n, c'est-à-dire à partir de Q_n. Pour $k \in \{1, \ldots, n\}$, la masse $M^o_{n,k}$ d'un organe o d'âge k au cycle de développement n est la somme des quantités de biomasse reçues depuis sa création :

$$M^o_{n,k} = \sum_{l=0}^{\min(k,T^o_{exp})-1} Al^{o,l}_{n-k+l}(Q_{n-k+l}, N_{0\to n-k+1+l}, \Theta_{fonc}), \qquad 1 \leq n \text{ et } 1 \leq k \leq n. \quad (2.4)$$

N.B. 2.11 L'équation précédente n'est pas définie pour les organes d'âge 0. Nous adoptons la convention qu'un organe d'âge 0 possède une masse nulle (aucune biomasse ne lui a encore été allouée). Ce n'est seulement qu'à la fin de son premier cycle de développement que celui-ci reçoit de la biomasse pour la première fois (voir la section 2.2.3).

2.2.3 Intégration dans le cycle de développement

Déroulement d'un cycle de développement

A chaque cycle de développement, trois phénomènes biophysiques se produisent : l'organogenèse, la photosynthèse et l'allocation. L'objectif de ce paragraphe est de présenter le déroulement du cycle de développement de façon générale et de préciser entre autre comment ces phénomènes sont liés les uns aux autres.

N.B. 2.12 Du point de vue modélisation, nous supposons que la différenciation des bourgeons apicaux (si elle a lieu) s'effectue au niveau de l'étape d'organogenèse, juste après la création des nouveaux organes. Ce choix est cohérent avec les priorités établies pour les règles de production (voir la section 2.1.7).

Chaque cycle de développement $n \geq 1$ débute par la photosynthèse. La plante synthétise alors une quantité Q_n de biomasse. Puis, le cycle se poursuit avec l'étape d'organogenèse. Les bourgeons créent de nouveaux phytomères suivant un ensemble de règles de production. Les organes ainsi créés possèdent un âge 0. La biomasse Q_n est ensuite répartie entre les différents organes en expansion (incluant les organes d'âge 0). Au début du cycle $n + 1$, tous les organes voient leur âge augmenter de un.

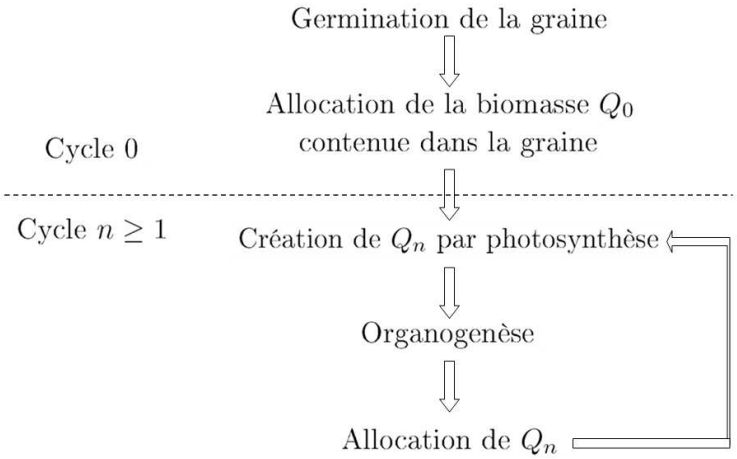

FIG. 2.9 – Principales étapes de la croissance.

Le cycle 0 est légèrement différent puisque la biomasse n'est pas créée par photosynthèse mais provient des réserves contenues dans la graine. Durant ce cycle, l'organogenèse est le résultat de la germination de la graine. Le déroulement de la croissance de la plante est résumée dans la figure 2.9.

Prenons, par exemple, la plante dont l'évolution est donnée par la figure 2.7. Supposons que $T^o_{exp} = 2$, pour tout $o \in \mathcal{O}$. Au début du cycle de développement 1, la plante est constituée d'un entrenœud et d'un bourgeon apical de CP 1, d'un bourgeon axillaire de CP 2 et d'une feuille. Tous ces organes sont âgés d'un cycle. Toujours au début de ce même cycle, elle produit une quantité Q_1 de biomasse par photosynthèse. Ensuite, le bourgeon de CP 1 va créer un phytomère de CP 1 et le bourgeon de CP 2 un phytomère de CP 2 (ces organes ont un âge de 0 cycle). Ces deux phytomères possèdent alors une masse nulle et ne sont pas encore visibles sur la plante. Enfin, la biomasse Q_1 est distribuée à l'ensemble des organes en expansion (dans ce cas, les organes d'âge 0 et 1). Au début du cycle 2, les phytomères créés au cycle 1 apparaissent physiquement (puisqu'ils ont une masse) et leur âge passe de 0 à 1 cycle.

Equation récurrente de photosynthèse

A partir des équations écrites dans la section 2.2.2, il est possible d'écrire l'équation de production (2.3) de façon récurrente. En effet, nous allons montrer que, à un cycle n donné, la quantité de biomasse Q_n créée est une fonction de Q_k avec $k \leq n-1$. La photosynthèse au cycle n dépend en fait des surfaces des feuilles actives durant ce même cycle. On qualifie d'active toute feuille susceptible d'intercepter la lumière du soleil durant la photosynthèse. Chez certaines plantes comme la betterave, il est possible que les feuilles soient actives pendant plusieurs cycles de développement avant de mourir. Dans le métamodèle $\mathcal{M}$, la sénescence des feuilles est prise en compte en introduisant un temps d'activité. Soit alors $T_{act} \in \mathbb{N}^*$ le nombre de cycles de développement durant lesquels une feuille est active (pour les arbres en zone tempérée, $T_{act} = 1$: les feuilles apparaissent au printemps et tombent en automne). Pour $n \in \mathbb{N}^*$ et $k \in \{1, \ldots, T_{act}\}$, soit $S_{n,k}$ la surface d'une feuille active d'âge k au cycle de développement n. La biomasse Q_n créée est alors une fonction de :

– $S_{n,1},\ldots, S_{n,T_{act}}$ si $n \geq T_{act}$;
– $S_{n,1},\ldots, S_{n,n}$ si $n < T_{act}$.

De façon générale, on a donc plus précisément :

$$Q_n = \Phi_n(S_{n,1}, \ldots, S_{n,\min(n,T_{act})}, N_{0\to n}, E_n, \Theta_{fonc}), \qquad n \geq 1. \tag{2.5}$$

N.B. 2.13 Une feuille ne peut être active que si elle a une surface et donc une masse. Ainsi, si une feuille est créée au cycle de développement n, elle ne sera pas active durant ce même cycle puisque sa masse est nulle. Par contre, elle le sera au cycle $n+1$ et ceci jusqu'au cycle $n + T_{max}$.

La surface d'une feuille est reliée à la masse de son limbe grâce à une fonction de densité $d_{n,k}$ qui peut dépendre de l'âge k de la feuille et du cycle de développement n en cours :

$$S_{n,k} = d_{n,k}(M^l_{n,k}) \qquad 1 \leq n \ \text{ et } \ 1 \leq k \leq n \tag{2.6}$$

avec $l \in \mathcal{O}$ le symbole désignant le limbe d'une feuille. La biomasse créée Q_n est donc une fonction de $M^l_{n,k}$ avec $k \in \{1, \ldots, \min(n, T_{act})\}$. Or, d'après l'équation (2.4), $M^l_{n,k}$ est une fonction de Q_{n-k+m} avec $m \in \{0, \ldots, \min(k, T^l_{exp}) - 1\}$. Dans ce cas, l'équation de production peut se réécrire comme une fonction, que nous noterons encore Φ_n, des biomasses créées aux cycles $k \in \{n - \min(n, T_{act}), \ldots, n-1\}$:

$$Q_n = \Phi_n(Q_{(n-T_{act})^+}, \ldots, Q_{n-1}, N_{0\to n}, E_n, \Theta_{fonc}), \qquad n \geq 1. \tag{2.7}$$

avec $(n - T_{act})^+ \stackrel{\text{def}}{=} \max(n - T_{act}, 0) = n - \min(n, T_{act})$. Cette équation joue un rôle capital pour l'estimation du jeu de paramètres Θ_{fonc} car elle permet d'écrire le système dynamique sous la forme d'un modèle de Markov caché (voir le chapitre 7).

N.B. 2.14 Les masses $M^l_{n,k}$ avec $k \in \{1, \ldots, \min(n, T_{act})\}$ dépendent aussi de $N_{0\to n-k+m}$ avec $m \in \{1, \ldots, \min(k, T^l_{exp})\}$. Or, d'après l'équation (2.2), toute l'information contenue dans $N_{0\to m}$ l'est aussi dans $N_{0\to m'}$ si $m \leq m'$. Ainsi, nous retrouvons que Q_n ne dépend que de $N_{0\to n}$.

2.3 Exemple du modèle GreenLab

Nous présentons dans cette section les principales caractéristiques du modèle de croissance de plante GreenLab. Une description complète du modèle est donnée dans Mathieu (2006) et Letort (2008) (en particulier, les calculs menant aux équations présentées dans ce qui suit). GreenLab appartient à la classe des modèles de croissance de plantes définie par le métamodèle $\mathcal{M}$. Il sera choisi comme exemple d'application par la suite.

Le modèle GreenLab est un modèle générique de croissance de plante (c'est-à-dire qu'il n'est pas dédié à une espèce de plante en particulier). Il peut en effet aussi bien être utilisé pour des petites plantes comme l'*Arabidopsis* (une vingtaine de centimètres, voir Christophe et al. (2008)) que pour des grands arbres comme le hêtre (une dizaine de mêtres, voir Letort et al. (2008a)). Dans cette thèse, nous nous intéressons à la

croissance de la plante isolée : l'interaction entre la croissance de la plante et celle de ces voisines n'est pas prise en compte. Le passage au peuplement a été déjà traité en partie (voir Cournède et al. (2008) pour les questions de compétition ou encore Le Chevalier (2010) pour les paysages fonctionnels) et fait encore l'objet de recherches. Les modèles de peuplement intègrent en outre la compétition concernant le partage des ressources naturelles (lumière, eau, ...). Cela ne sera pas pris en compte ici. De même, les équations présentées par la suite sont valables pour un modèle GreenLab dans lequel la croissance secondaire ne dépend pas de la topologie. Nous ne considérons donc pas le cas où l'allocation à la croissance secondaire se fait selon le principe du « pipe model », dans lequel la croissance en épaisseur d'un entrenœud dépend de la surface des feuilles actives situées au dessus de lui dans la topologie de la plante (voir Shinozaki et al. (1964) et Letort (2008)).

2.3.1 Description des processus biophysiques

Nous reprenons chacun des processus biophysiques du fonctionnement détaillés précédemment et les décrivons dans le cadre du modèle GreenLab. Le modèle de développement n'est pas expliqué car il est pratiquement identique au modèle de la section 2.1 à quelques mots de vocabulaire près.

Photosynthèse

L'un des objectifs de GreenLab est de proposer un modèle de croissance prenant en compte les éléments écophysiologiques essentiels tout en conservant une formulation mathématique simple. Dans cette optique, le processus de photosynthèse se résume à une seule équation intégrant les effets cumulés sur un cycle de développement de plusieurs phénomènes biophysiques tels que le transport de l'eau ou encore l'impact des conditions environnementales (température, ensoleillement, ...). En outre, dans un souci de simplicité, le modèle GreenLab n'intègre pas dans ses équations la position spatiale des organes et notamment celle des feuilles pour la photosynthèse. Ainsi, l'interception lumineuse n'est pas calculée de façon exacte mais est modélisée par une loi de type Beer-Lambert (voir Marcelis et al. (1998)). La surface foliaire interceptant la lumière ($= SIL$) est alors donnée par :

$$SIL = S_p \left(1 - exp\left(-\frac{k_b S_n^{tot}}{S_p}\right)\right) \qquad n \geq 1. \tag{2.8}$$

avec S_n^{tot} la somme des surfaces des feuilles participant activement à la photosynthèse :

$$S_n^{tot} = \sum_{k=1}^{\min(n,T_{act})} \left(N_{n+1-k}^l - N_{n-k}^l\right) S_{n,k} \qquad n \geq 1. \tag{2.9}$$

avec $l \in \mathcal{O}$ le symbole désignant le limbe d'une feuille. k_b est le coefficient d'extinction de la loi de Beer-Lambert. Celui-ci est lié, botaniquement parlant, à l'inclinaison des feuilles et à la direction moyenne du rayonnement lumineux. S_p est une surface foliaire caractéristique. Elle est spécifique à un modèle de plante donné. Elle peut parfois s'interpréter comme la projection du couvert végétal sur le sol suivant la direction moyenne du

rayonnement lumineux. Bien que les hypothèses validant l'utilisation de la loi de Beer-Lambert soient rarement vérifiées dans la réalité (feuilles infiniment petites et réparties aléatoirement de façon uniforme sur la plante), l'approximation faite par l'équation (2.8) sur la surface foliaire interceptant la lumière reste bonne en pratique.

La quantité Q_n de biomasse créée au cours du cycle de développement n est proportionnelle à la SIL. L'équation de production (2.3) est donnée de façon générale par :

$$Q_n = E_n \mu S_p \left(1 - exp\left(-\frac{k_b S_n^{tot}}{S_p}\right)\right) \qquad n \geq 1. \tag{2.10}$$

μ est un paramètre permettant de calibrer au mieux la production photosynthétique pour un modèle de plante donné. E_n traduit toujours l'impact de l'environnement au cycle n. Les données expérimentales utilisées sont des moyennes observées sur un cycle de développement. Lorsque seul l'ensoleillement est pris en compte (dans un soucis de simplification du modèle ou par manque d'autres données environnementales), l'équation (2.10) est équivalente à l'équation de production de Monteith (voir Monteith (1977)), initialement établie dans le cadre des cultures en champs mais adaptée pour l'étude de la plante isolée :

$$Q_n = PAR_n \, RUE \, S_p \left(1 - exp\left(-\frac{k_b S_n^{tot}}{S_p}\right)\right) \qquad n \geq 1. \tag{2.11}$$

PAR signifie Photosynthetically Active Radiation. PAR_n représente l'énergie lumineuse totale reçue par unité de surface si toute l'énergie lumineuse incidente est absorbée par le couvert végétal dans la gamme de longueur d'onde utile pour la photosynthèse. RUE signifie Radiation Use Efficiency. Il s'agit d'un coefficient de conversion de l'intensité lumineuse reçue en biomasse produite. Ainsi, $PAR_n \, RUE$ donne la quantité de biomasse produite au cycle n par mètre carré de couvert végétal. En multipliant par la SIL, nous obtenons bien Q_n.

N.B. 2.15 Le vecteur de paramètres endogènes Θ_{fonc} contient μ (respectivement RUE), k_b et S_p si l'équation de photosynthèse utilisée est (2.10) (respectivement (2.11)).

Allocation

L'allocation se fait selon un modèle sources-puits. Les organes sources sont ceux qui participent à la production de biomasse (les feuilles dans la majotité des cas). Les organes puits consomment la biomasse. Il s'agit donc de l'ensemble des organes en expansion. Supposons qu'une quantité Q_n de biomasse ait été produite par photosynthèse au cours du cycle n. Chacun des organes en expansion reçoit alors un certain pourcentage de Q_n qui correspond à sa demande relative en biomasse (modèle d'allocation proportionnelle, Warren-Wilson (1972)). C'est l'hypothèse du pool commun. La demande d'un organe représente sa capacité à attirer les assimilats créés lors de la photosynthèse. Elle est mesurée par une fonction puits. Plus le puits d'un organe est élevé par rapport aux autres, plus celui-ci reçoit de biomasse. Nous supposons que le puits ne dépend que du type $o \in \mathcal{O}$ de l'organe ainsi que de son âge. En outre, il ne dépend pas de sa position

sur la plante. Le puits est alors donné par la formule suivante :

$$p_o(k) = P_o N_o \left(\frac{k+0.5}{T^o_{exp}}\right)^{\alpha_o - 1} \left(1 - \frac{k+0.5}{T^o_{exp}}\right)^{\beta_o - 1} \qquad k \in \{0, \ldots, T^o_{exp} - 1\}. \tag{2.12}$$

α_o et β_o sont des paramètres permettant de contrôler la forme des puits (voir le *Nota Bene* 2.16 ci-dessous). P_o désigne la force de puits d'un organe. N_o est une constante de normalisation choisie de sorte que $\max_{k \in \{0,\ldots,T^o_{exp}-1\}} p_o(k) = P_o$. P_o, α_o et β_o sont des paramètres inconnus *a priori* et font partie de Θ_{fonc}.

N.B. 2.16 La fonction puits p_o est en fait proportionnelle à la densité d'une loi bêta de paramètres α_o et β_o. Ce choix se justifie par la richesse des types de courbes obtenues en ne modifiant que les paramètres α_o et β_o (voir la figure 2.10). Il est ainsi possible d'obtenir toutes sortes de fonctions puits en ne calibrant seulement que deux paramètres ce qui est bénéfique pour l'estimation des paramètres du modèle (voir Yin et al. (2003)).

N.B. 2.17 Si $T^o_{exp} \leq 2$, d'autres fonctions puits peuvent être proposées afin d'éviter une surparamétrisation du modèle (voir la section 7.3).

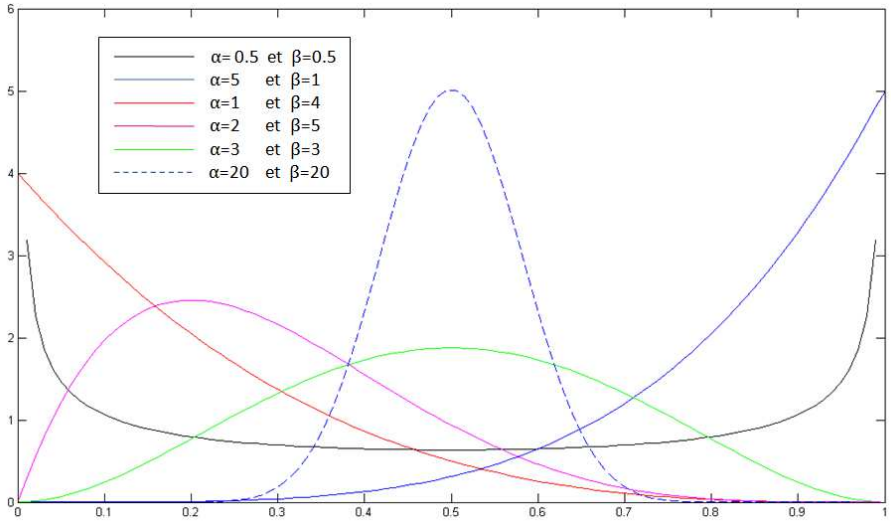

FIG. 2.10 – Les différentes formes possibles de la densité de la loi beta de paramètres (α, β)

La demande relative en biomasse d'un organe est le rapport entre la demande d'un organe et la demande totale D_n des organes en expansion au cycle de développement n avec :

$$D_n = \sum_{o \in \mathcal{O}} \sum_{k=0}^{\min(T^o_{exp}-1, n)} \left(N^o_{n+1-k} - N^o_{n-k}\right) p_o(k) \qquad n \geq 0. \tag{2.13}$$

La quantité de biomasse $Al^{o,l}_n(Q_n, N_{0\to n+1}, \Theta_{fonc})$ allouée à un organe o d'âge l au cycle de développement n est alors donnée par :

$$Al^{o,l}_n(Q_n, N_{0\to n+1}, \Theta_{fonc}) = \frac{p_o(l)}{D_n} Q_n = p_o(l) \frac{Q_n}{D_n} \tag{2.14}$$

En reprenant l'équation (2.4), la masse $M^o_{n,k}$ d'un organe o d'âge k et créé au cycle n est donnée par :

$$M^o_{n,k} = \sum_{l=0}^{\min(k,T^o_{exp})-1} p_o(l)\frac{Q_{n-k+l}}{D_{n-k+l}} \qquad 1 \leq n \text{ et } 1 \leq k \leq n. \tag{2.15}$$

N.B. 2.18 En général, dans GreenLab, la fonction de densité (voir l'équation (2.6)) ne dépend pas de l'âge de la feuille ni du cycle de développement en cours. Elle est donnée par :

$$S_{n,k} = M^l_{n,k}/e \qquad 1 \leq n \text{ et } 1 \leq k \leq n$$

avec e la densité surfacique des feuilles. En combinant les équations (2.10), (2.9), (2.13) et (2.15), nous obtenons l'équation récurrente de photosynthèse (2.7).

N.B. 2.19 L'hypothèse d'un pool commun a été introduite dans GreenLab afin de simplifier le modèle d'allocation. Une telle hypothèse suggère que la quantité de biomasse reçue par un organe ne dépend pas de sa position sur la plante. Cependant, celle-ci n'est pas toujours vérifiée (elle est valable par exemple pour la tomate, Heuvelink (1995), mais pas pour la vigne, Pallas et al. (2009)). Généralement, elle est valable pour les petites plantes mais est rarement vraie chez les arbres pour lesquels la répartition de la biomasse se fait plus localement (phénomène lié au problème de transport des assimilats).

2.3.2 Algorithme de croissance

Soit $T_{sim} \in \mathbb{N}^*$ le temps maximal de simulation (en cycle de développement). Le modèle de croissance GreenLab peut se résumer par l'algorithme suivant :

Algorithme 1 : Modèle de croissance GreenLab

- **Initialisation** : cycle 0

$ Etat initial de la plante :
- Poser $N^a_0 = \delta_{b_1}(a)$, $a \in \mathcal{B} \cup \mathcal{O}$.
- Poser $N_0 = (\ldots, N^a_0, \ldots)_{b \in \mathcal{B}\cup\mathcal{O}}$.
- Poser Q_0 : biomasse contenue dans la graine.

$ organogenèse :
- Calculer N_1 à partir de N_0 et des règles de production.

$ Allocation :
- Calcul de la demande totale au cycle 0 :

$$D_0 = \sum_{o \in \mathcal{O}} N^o_1 p_o(0).$$

- Pour $o \in \mathcal{O}$:
  - calculer la masse des organes au début du cycle 1 :

$$M^o_{0,1} = p_o(0)\frac{Q_0}{D_0}.$$

• **Itération** :
- Pour $n = 1, \ldots, T_{sim}$, faire :
  $ Photosynthèse du cycle n :
  - Calculer la surface foliaire active totale :

$$S_n^{tot} = \sum_{k=1}^{\min(n,T_{act})} \left(N_{n+1-k}^l - N_{n-k}^l\right) S_{n,k} = \sum_{k=1}^{\min(n,T_{act})} \left(N_{n+1-k}^l - N_{n-k}^l\right) \frac{M_{n,k}^l}{e}.$$

  - Calculer la biomasse produite :

$$Q_n = E_n \mu S_p \left(1 - exp\left(-\frac{k_b S_n^{tot}}{S_p}\right)\right)$$

  $ organogenèse du cycle n :
  - Calculer N_{n+1} à partir de N_n et des règles de production.
  $ Allocation du cycle n :
  - Calculer la demande totale :

$$D_n = \sum_{o \in \mathcal{O}} \sum_{k=0}^{\min(T_{exp}^o - 1, n)} \left(N_{n+1-k}^o - N_{n-k}^o\right) p_o(k)$$

  - Pour $o \in \mathcal{O}$ et $k = 1, \ldots, n+1$, faire :
    - Calculer la masse au début du cycle $n+1$ d'un organe o d'âge k :

$$M_{n+1,k}^o = \sum_{l=0}^{\min(k, T_{exp}^o)-1} p_o(l) \frac{Q_{n+1-k+l}}{D_{n+1-k+l}}.$$

L'initialisation correspond au cycle 0 de la croissance. La plante est alors à l'état de graine. Le vecteur N_0 ne contient donc que des composantes nulles à l'exception de celle correspondant à un bourgeon de CP 1 qui vaut 1. La graine germine ensuite pour faire apparaître les premiers organes de la plante qui seront comptabilisés dans N_1 (les bourgeons qui seront actifs au cycle de développement 1 sont entre autre créés lors de cette étape). Les organes fraîchement créés reçoivent une fraction de la biomasse Q_0 contenue dans la graine correspondant à leur demande relative. Au début du cycle 1, ces organes deviennent visibles sur la plante et sont âgés d'un cycle. De façon générale, le cycle $n \geq 1$ débute par la photosynthèse. Cette étape commence par le calcul de la surface foliaire active totale S_n^{tot} qui s'exprime en fonction des masses des limbes calculées lors du cycle précédent. La quantité de biomasse Q_n est ensuite déterminée par l'équation (2.10). Puis, le cycle continue avec la création de nouveaux organes (comptabilisés dans N_{n+1}) par les bourgeons actifs de la plante (comptabilisés dans N_n). Enfin, l'étape d'allocation débute par le calcul de la demande totale de la plante. La biomasse Q_n est alors distribuée à l'ensemble des organes en expansion en fonction de leur demande relative. Le cycle se termine par le calcul des nouvelles masses des organes (masses visibles sur la plante au début du cycle suivant). L'âge de la plante (et de tous ses organes) augmente alors de 1.

2.3.3 Les différentes versions du modèle

Il existe plusieurs versions du modèle GreenLab différant chacune par le type d'organogenèse considéré et la présence ou non de rétroaction du fonctionnement de la plante sur son développement.

La première version de GreenLab , nommée GL1, se caractérise par une organogenèse déterministe et indépendante du fonctionnement de la plante (voir de Reffye et al. (2003)). Bien qu'étant la version la plus simple à ce jour, celle-ci trouve de nombreuses applications (le maïs, Guo et al. (2006), la betterave, Lemaire et al. (2008), la tomate, Dong et al. (2008) ou encore le coton, Dong et al. (2009)). La décomposition de la plante selon un ensemble de structures identiques (voir la méthode des sous-structures, Cournède et al. (2006)) permet un gain computationnel important (le temps de simulation passe d'exponentiel à polynomial).

La seconde version (GL2, voir Kang et al. (2008)) introduit un modèle d'organogenèse stochastique (voir le chapitre 3 pour plus de détails) toujours indépendant du fonctionnement. Ceci permet d'intégrer simplement dans le modèle d'organogenèse des phénomènes naturels imprévisibles (bourgeon mangé par les insectes,...) ou encore des processus de développement encore mal compris (nombre variable de phytomères engendrés par un bourgeon,...). Il est ainsi possible de simuler la variété architecturale au sein d'une même espèce de plante. Ce modèle s'applique à des plantes comme le pin sylvestre (variété mongolienne), Wang et al. (2010).

Le modèle GL3 revient à un modèle d'organogenèse déterministe mais introduit une rétroaction du fonctionnement sur le développement (voir Mathieu (2006) and Mathieu et al. (2009)). Ainsi, les règles de production associées au développement des bourgeons dépendent d'une variable rendant compte de la compétition trophique au sein de la plante. A un cycle de développement donné, cette variable est en fait le rapport entre la biomasse disponible et la demande totale des organes en expansion. Si le rapport offre sur demande est élevé, alors les bourgeons produiront de nouveaux phytomères. En revanche, dans le cas contraire, un grand nombre de bourgeons restera inactif durant le cycle en question et ne produira rien. Un tel modèle permet de prendre en compte l'impact de l'environnement sur le développement architectural de la plante. Ainsi, suivant les conditions environnementales considérées, une même plante possèdera des structures différentes. Cette version de GreenLab permet une meilleure compréhension de la croissance et gagne en popularité. Elle est appliquée entre autre sur la tomate (Kang et al. (2010)), le poivron (Ma et al.) et le hêtre (Letort et al. (2008a)).

Le modèle GL4 intègre l'organogenèse stochastique de GL2 et la rétroaction de GL3. Ainsi, les probabilités associées aux règles de production dépendent de la compétition trophique (c'est-à-dire de l'offre de biomasse sur la demande totale). Ce modèle est encore à l'étude et n'a pas pour le moment de forme définitive. Un premier essai a cependant été effectué sur la vigne, Pallas et al. (2009) et Pallas et al. (2010).

Un dernier modèle (GL5) est en cours de développement. Celui-ci se focalise plus sur la croissance des arbres et intègre notamment le polycyclisme, la préformation et la néoformation.

Chapitre 3

Modélisation stochastique du développement de la plante

Le développement des bourgeons peut être influencé par un certain nombre de facteurs qui affectent leur comportement. Ces facteurs peuvent être externes à la plante (bourgeons mangés par des insectes, incidents météorologiques, ...) ou internes (processus de développement dont la mécanique n'est pas entièrement comprise ou trop complexe) et sont souvent difficiles à prévoir ou mal compris. Afin d'étudier au mieux la croissance des plantes en question, il est bien de les prendre en compte et de les intégrer dans le modèle. Il en résulte que le comportement des bourgeons n'est plus certain et que plusieurs évolutions sont possibles lors de leur développement. Dans ce cas, une probabilité d'occurrence est associée à chacune de ces évolutions. Nous pouvons mettre en évidence deux types de processus aléatoires : ceux qui sont liés à l'organogenèse et ceux qui sont liés à la différenciation du bourgeon apical d'un axe (cf la section 2.1.4). Dans ce chapitre, nous présentons tout d'abord ces deux types de processus stochastiques (sections 3.1 et 3.1.3). Les processus ainsi présentés définissent un modèle de développement stochastique que nous appellerons $\mathcal{S}$ et qui vient compléter le modèle de développement de $\mathcal{M}$ (cf chapitre 2). Les sections suivantes sont consacrées à l'étude de $\mathcal{S}$. Cette étude est importante pour deux raisons. Tout d'abord, cela permet de comprendre, d'analyser et de critiquer le modèle ainsi défini en confrontant son comportement avec celui de vraies plantes. Ensuite, son étude est une étape prérequise pour l'estimation du vecteur de paramètres Θ_{fonc} lié au fonctionnement du métamodèle $\mathcal{M}$ (voir le chapitre 7). Dans un premier temps, nous mettons en évidence les processus de branchement sous-jacents à $\mathcal{S}$ (section 3.3). La section suivante introduit les fonctions génératrices permettant de calculer des quantités d'intérêt concernant la composition de la plante. Enfin, la dernière section présente une courte étude sur la finitude de la croissance. Les variables (ou vecteurs) aléatoires présentées par la suite sont définies sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.

3.1 Processus aléatoires de l'organogenèse

3.1.1 Description

Nous présentons dans cette section l'ensemble des processus aléatoires intervenant dans l'organogenèse. Etant donné le nombre important de phénomènes à prendre en compte, ceux-ci sont résumés par quatre principaux processus stochastiques (généralisant ceux présentés dans de Reffye (1979) et Kang et al. (2008)) :

- probabilité de survie $p_s(i)$, $i \in \{1, \ldots, CP_m\}$: au début de l'organogenèse, un bourgeon de CP i peut mourir avec une probabilité $1-p_s(i)$.
- probabilité d'activité $p_a(i)$, $i \in \{1, \ldots, CP_m\}$: si un bourgeon de CP i ne meurt pas, il se peut que celui-ci soit dormant (c'est-à-dire qu'il ne produit aucun nouvel organe durant l'organogenèse) avec une probabilité $1-p_a(i)$. Dans le cas contraire, le bourgeon est dit actif.
- distribution du nombre d'entrenœuds produits $(p_{e_i}(k_0, k_1, \ldots, k_{CP_m}))_{(k_0,\ldots,k_{CP_m})\in\mathbb{N}^{CP_m+1}}$, $i \in \{1, \ldots, CP_m\}$: lorsqu'un bourgeon de classe CP i est actif, celui-ci produit, pour chaque $j \in \{1, \ldots, CP_m\}$, k_j entrenœuds de CP i portant chacun potentiellement des bourgeons axillaires de CP j et k_0 entrenœuds de CP i ne portant aucun bourgeon axillaire avec une probabilité $p_{e_i}(k_0, k_1, \ldots, k_{CP_m})$. Le nombre total de phytomères créés par ce même bourgeon vaut donc $\sum_{j=0}^{CP_m} k_j$. Etant donné les règles botaniques énoncées dans la section 2.1.2, $p_{e_i}(k_0, k_1, \ldots, k_{CP_m}) = 0$ s'il existe $0 < j < i$ tel que $k_j > 0$ puisqu'un entrenœud d'une CP donnée ne peut porter que des axes latéraux de CP supérieure ou égale. Bien que cette distribution soit sur $\mathbb{N}^{CP_m+1}$, pour chaque $j \in \{0, \ldots, CP_m\}$, il existe un rang que nous notons $Ent^{i,j}_{max} \in \mathbb{N}$ tel que si $k_j > Ent^{i,j}_{max}$ alors $p_{e_i}(k_0, \ldots, k_{CP_m}) = 0$. $Ent^{i,j}_{max}$ est une donnée botanique propre à chaque plante.
- distribution du nombre de bourgeons axillaires portés par un entrenœud $\left(p^{e_i}_{b_j}(k)\right)_{k\in\mathbb{N}}$, $(i, j) \in \{1, \ldots, CP_m\}^2$: lorsqu'un entrenœud de CP i portant des bourgeons axillaires de CP j est créé, le nombre de bourgeons latéraux est égal à k avec une probabilité $p^{e_i}_{b_j}(k)$. Nous supposons que le nombre de bourgeons créés vaut au moins 1. Dans ce cas, $p^{e_i}_{b_j}(0) = 0$. Les variables aléatoires donnant le nombre de bourgeons latéraux portés par un entrenœud sont supposées indépendantes. Bien que la distribution soit sur $\mathbb{N}$, il existe un rang que nous notons $Bou^{i,j}_{max} \in \mathbb{N}$ tel que si $k > Bou^{i,j}_{max}$ alors $p^{e_i}_{b_j}(k) = 0$. $Bou^{i,j}_{max}$ est également une donnée botanique propre à chaque plante.

Nous rappelons que les bourgeons se comportent indépendamment les uns des autres. Chaque bourgeon vivant au début d'un cycle de développement donné est soumis aux quatre processus précédents selon un ordre de priorité établi. Ainsi, à chaque étape d'organogenèse, la première expérience aléatoire à laquelle sont soumis les bourgeons est la question de la survie. Viennent ensuite successivement l'activité, le nombre d'entrenœuds créés par le bourgeon et enfin le nombre de bourgeons axillaires portés par chaque entrenœud.

L'ensemble des probabilités intervenant dans le modèle d'organogenèse stochastique peut être regroupé sous la forme d'un vecteur Θ_{org} défini de la façon suivante :

$$\Theta_{org} = \{p_s(i)\}_i \cup \{p_a(i)\}_i \cup \{p_{e_i}(k_0, \ldots, k_{CP_m})\}_{i,k_0,\ldots,k_{CP_m}} \cup \{p_{b_j}^{e_i}(k)\}_{i,j,k}$$

avec $i \in \{1, \ldots, CP_m\}$, $j \in \{1, \ldots, CP_m\}$, $k_l \in \{0, \ldots, Ent_{max}^{i,l}\}$ pour tout $l \in \{0, \ldots, CP_m\}$ et $k \in \{0, \ldots, Bou_{max}^{i,j}\}$. Dans le chapitre 5, nous proposons une méthode permettant d'estimer Θ_{org} à partir de données expérimentales sur un ensemble de vraies plantes.

N.B. 3.1 En réalité, le nombre de bourgeons axillaires portés par un entrenœud et créés à un cycle n est une donnée botanique constante (propre à la plante étudiée) qui ne dépend que de la CP de l'entrenœud auquel ils sont rattachés. Deux entrenœuds d'un même type porteront donc le même nombre de bourgeons axillaires. Cependant, ces bourgeons ne donneront pas tous un axe latéral (ils sont alors qualifiés de potentiels). Ce n'est que lors de l'organogenèse du cycle de développement suivant (c'est-à-dire le cycle $n+1$) que le nombre définitif d'axes latéraux sera fixé (dans ce cas, il est possible que le nombre d'axes latéraux diffère pour des entrenœuds ayant même CP). Ce phénomène naturel est modélisé par un processus aléatoire caractérisé par la probabilité de branchement d'un bourgeon potentiel. Dans le cadre du modèle d'organogenèse présenté dans cette thèse, ce processus ne dépend que de la CP des bourgeons potentiels (en particulier, il ne dépend pas du cycle de développement considéré). Afin de simplifier le modèle d'organogenèse, nous supposerons que, pour des bourgeons potentiels créés au cycle n, l'expérience aléatoire liée au branchement des bourgeons potentiels se déroule également au cycle n (et non au cycle $n+1$ comme il le devrait) et est caractérisée par les distributions $\left(p_{b_j}^{e_i}(k)\right)_{k\in\mathbb{N}}$, $(i,j) \in \{1, \ldots, CP_m\}^2$. Cette simplification n'affecte en rien l'étude stochastique des plantes étudiées dans la thèse. Cependant, elle ne serait plus valable si nous choisissions de faire dépendre les probabilités de branchement du cycle de développement et notamment de l'état de la plante (ce qui est le cas pour des modèles comme GreenLab 4).

3.1.2 Modèle d'organogenèse de GreenLab2

Le modèle GreenLab 2 (cf la section 2.3.3) est un cas particulier du modèle d'organogenèse stochastique décrit dans la section précédente (voir Kang et al. (2008) pour une description complète du modèle d'organogenèse de GreenLab 2). Les probabilités de survie $p_s(i)$ et d'activité $p_a(i)$ sont définies de la même façon. La distribution $(p_{e_i}(k_0, \ldots, k_{CP_m}))_{(k_0,\ldots,k_{CP_m})\in\mathbb{N}^{CP_m+1}}$ se décompose en un produit de distribution $\left(p_{e_{i,j}}(k_j)\right)_{(k_j)\in\mathbb{N}}$ de la façon suivante :

$$p_{e_i}(k_0, \ldots, k_{CP_m}) = \prod_{j=1}^{CP_m} p_{e_{i,j}}(k_j)$$

avec $p_{e_{i,j}}(k_j)$ la probabilité pour un bourgeon de CP i de produire k_j entrenœuds de CP i pouvant potentiellement porter des bourgeons latéraux de CP j. La distribution $\left(p_{e_{i,j}}(k_j)\right)_{(k_j)\in\mathbb{N}}$ est celle d'une loi binomiale de paramètres $(Ent_{max}^{i,j}, p_m(i,j))$ avec $p_m(i,j)$

la probabilité d'apparition d'un entrenœud de CP i portant potentiellement des bourgeons latéraux de CP j. De même, la distribution du nombre de bourgeons axillaires de CP j portés par un entrenœud de CP i $\left(p_{b_j}^{e_i}(k)\right)_{k\in\mathbb{N}}$ est celle d'une loi binomiale de paramètres $(Bou_{max}^{i,j}, p_b(i,j))$ avec $p_b(i,j)$ la probabilité de branchement d'un bourgeon de CP j qui est porté latéralement par un entrenœud de CP i. Dans le cas du modèle GreenLab 2, nous noterons Θ_{org}^{GL} le vecteur contenant toutes les probabilités du modèle d'organogenèse associé.

3.1.3 Représentation à l'aide d'automates stochastiques

Tout comme pour le cas déterministe, les règles de production du modèle d'organogenèse stochastique associé à $\mathcal{S}$ peuvent également être représentées par un ensemble d'automates. Ceux-ci donnent les règles d'évolution pour chaque type de bourgeon au cours d'une étape d'organogenèse. Les probabilités d'évolution d'un bourgeon vers une structure donnée sont indiquées au-dessus des structures correspondantes (cf figure 3.1).

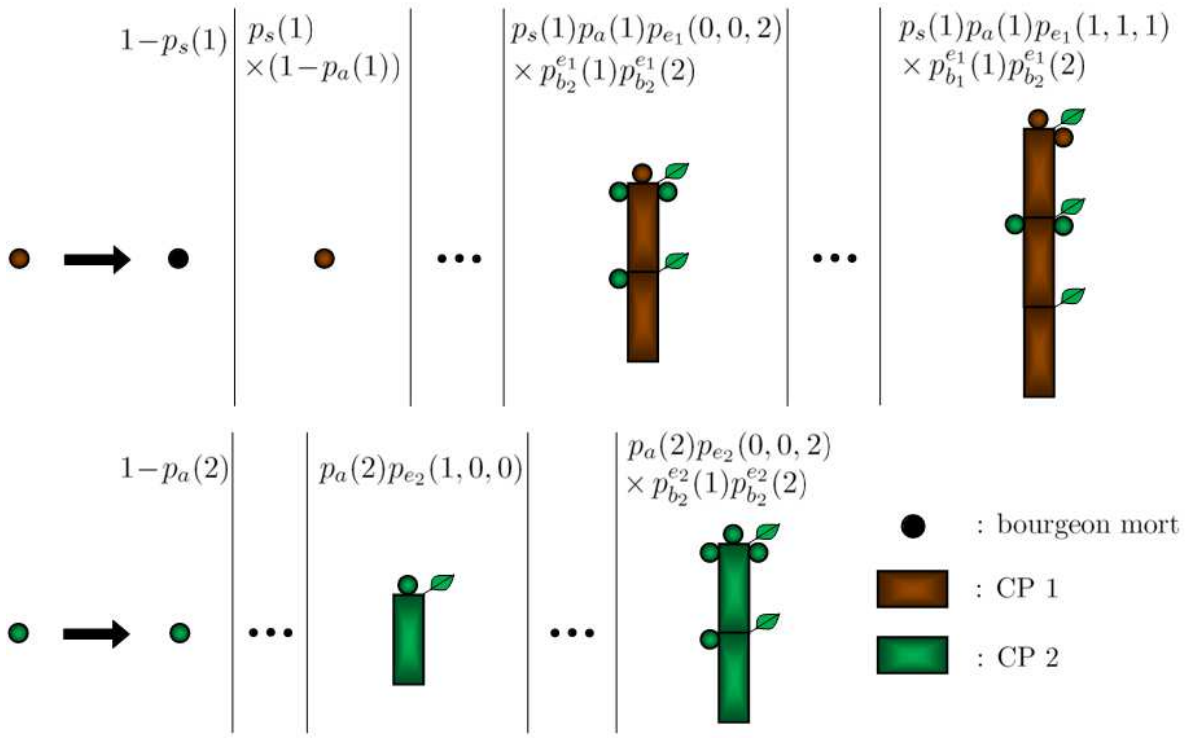

FIG. 3.1 – Exemple d'automates stochastiques avec $CP_m = 2$. Concernant l'automate du bas, la probabilité de survie $p_s(2)$ vaut 1.

3.2 Modélisation stochastique de la différenciation

Cette section présente nos travaux concernant la modélisation stochastique de la différenciation (cf Loi and Cournède (2008)). La différenciation (ou mutation) est le

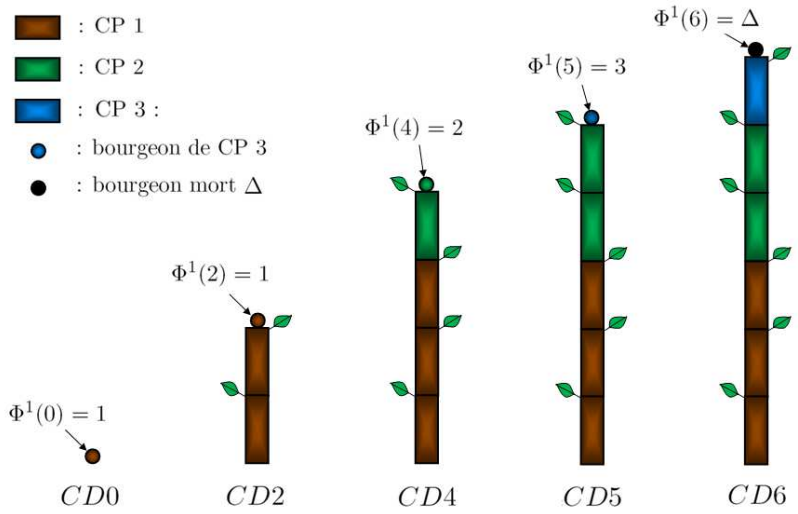

FIG. 3.2 – Processus de différenciation $\Phi^\beta(k)$. Dans l'exemple ci-dessus, $\beta = 1$. Δ correspond à l'état final de différenciation (ici la mort du bourgeon apical). CP = Classe Physiologique. CD = Cycle de Développement.

phénomène biophysique selon lequel le bourgeon apical d'un axe change de nature botanique sous l'effet du vieillissement (voir Barthélémy and Caraglio (2007) et la section 2.1.4). Elle se traduit par un changement de CP du bourgeon apical (vers une classe supérieure), le stade ultime étant la mort du bourgeon ou sa transformation en fleur. Soit $\Phi^\beta(k)$ la classe physiologique du bourgeon apical d'un axe d'âge k et ayant débuté par un bourgeon de CP β (c'est-à-dire $\Phi^\beta(0) = \beta$).

Etudier la differenciation revient donc à étudier le processus à temps discret $(\Phi^\beta(k))_{k\in\mathbb{N}}$ avec $\beta \in \{1, \ldots, CP_m\}$. Nous choisissons de modéliser $(\Phi^\beta(k))_{k\in\mathbb{N}}$ par un processus markovien de saut (voir Pardoux (2007)) que nous noterons $(\phi^\beta(t))_{t\in\mathbb{R}^+}$ de sorte que les deux processus coïncident sur les valeurs entières :

$$\forall k \in \mathbb{N}, \qquad \Phi^\beta(k) = \phi^\beta(k).$$

Ce choix se justifie principalement par deux points :

- le changement de classe physiologique d'un bourgeon apical est dû au vieillissement de l'axe auquel il appartient. Il est donc raisonnable de modéliser le temps de séjour d'un bourgeon apical dans une classe physiologique j donnée par une loi exponentielle de paramètre $\lambda_j > 0$ avec λ_j l'inverse du temps de séjour moyen dans la classe j (cette loi étant habituellement utilisée pour modéliser la durée de vie de certains phénomènes biologiques ou industriels notamment en théorie de la fiabilité, Bouleau (2002)).
- bien que travaillant avec une organogenèse discrète, il est préférable de modéliser la différenciation par un processus à temps continu pour des raisons d'estimation paramétrique (voir le chapitre 5). En effet, l'estimation des paramètres du processus passe par la mesure, à partir de véritables plantes, des temps de séjour moyens d'un bourgeon apical dans une classe donnée. Ces temps de séjour moyens étant

des nombres décimaux, modéliser la différenciation par un processus à temps discret nécessiterait de tronquer les valeurs mesurées sur le terrain. Ceci entraînerait une perte d'information sur le processus qui pourrait nuire à la qualité du modèle de croissance.

L'espace d'état $\mathcal{E}$ du processus markovien de saut est composé de toutes les classes physiologiques ainsi que d'un état final noté Δ représentant la mort du bourgeon ou sa transformation en fleur :

$$\mathcal{E} = \{1, \ldots, CP_m, \Delta\}.$$

Les états 1,...,CP_m sont tous transitoires. Seul Δ est absorbant. Soient $\{q_{i,j}, \quad (i,j) \in \Delta^2\}$ les probabilités de transition associées à la chaîne de Markov sous-jacente à $(\phi^\beta(t))_{t\in\mathbb{R}^+}$. $q_{i,j}$ représente la probabilité que le bourgeon apical passe de l'état i à l'état j au moment où celui-ci se différencie (ce qui correspond à un saut du processus). Etant donné qu'un bourgeon ne peut se différencier que vers une CP supérieure, nous avons $q_{i,j} = 0$ si $j \leq i$. Le générateur infinitésimal Q s'écrit donc :

$$Q = \left(\begin{array}{cccc|c} -\lambda_1 & \lambda_1 q_{1,2} & \ldots & \lambda_1 q_{1,CP_m} & \lambda_1 q_{1,\Delta} \\ 0 & -\lambda_2 & \ldots & \lambda_2 q_{2,CP_m} & \lambda_2 q_{2,\Delta} \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & -\lambda_{CP_m} & \lambda_{CP_m} \\ \hline 0 & 0 & \ldots & 0 & 0 \end{array}\right) = \left(\begin{array}{c|c} A & -A^t \mathbf{1}_{CP_m} \\ \hline \mathbf{0}_{CP_m} & 0 \end{array}\right)$$

avec A la sous-matrice de taille $CP_m \times CP_m$ du coin supérieur gauche et $\mathbf{0}_{CP_m}$ (resp. $\mathbf{1}_{CP_m}$) le vecteur ligne de taille CP_m dont toutes les composantes valent 0 (resp. 1). Comme $\phi^\beta(0) = \beta$, la distribution initiale du processus markovien de saut est donnée par la mesure de Dirac centrée en β : δ_β. Pour la suite, nous posons $\alpha_\beta = (\mathbf{0}_{\beta-1}, \delta_\beta, \mathbf{0}_{CP_m-\beta})$.

L'étude probabiliste de la différenciation se fait par l'étude des temps d'atteinte d'une classe physiologique donnée. Ces derniers sont étroitement liés à la théorie des phase-types (voir l'annexe A.2 concernant les principaux concepts). Les phase-types ont tout d'abord été développées dans les années 70 par Neuts (Neuts (1975)) avec de nombreuses applications en théorie de la fiabilité (Neuts (1981) et Assaf and Levikson (1982)). Par la suite, elles sont devenues un centre d'intérêt probabiliste qui a connu de nombreux développements mathématiques (Assaf et al. (1984), Kulkarni (1989) et plus récemment Goff (2005)). L'intérêt des phase-types réside dans le fait que de nombreux outils probabilistes peuvent être exprimés explicitement (notamment la fonction de répartition, voir l'annexe A.2) ce qui sera utile par la suite pour le calcul des fonctions génératrices ou encore l'étude de la finitude de la croissance (section 3.5). Soient $\{\Gamma_k\}_{k\in\{1,\ldots,CP_m+1\}}$ une famille de sous-ensembles de $\mathcal{E}$ définie de la façon suivante :

$$\Gamma_k = \begin{cases} \bigcup_{j=k}^{CP_m} \{j\} \cup \{\Delta\} & \text{si } k \in \{1, ..., CP_m\} \\ \{\Delta\} & \text{si } k = CP_m + 1 \end{cases}$$

Etant donné que $q_{i,j} = 0$ si $j \leq i$, les ensembles $\{\Gamma_k\}_{k\in\{1,\ldots,CP_m+1\}}$ sont bien stochastiquement fermés (voir l'annexe A.2 et la figure 3.3).

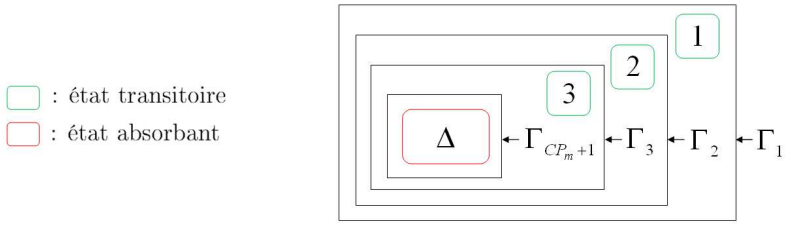

FIG. 3.3 – Représentation des ensembles Γ_k pour $CP_m = 3$.

Soit alors T_k^β le temps d'atteinte de Γ_k par ϕ^β :

$$T_k^\beta = \inf\{t \in \mathbb{R}^+, \phi^\beta(t) \in \Gamma_k\}.$$

Nous avons ainsi :

Proposition 3.2.1 *Pour tout $\beta \in \{1, \ldots, CP_m\}$, $(T^\beta_{\beta+1}, ..., T^\beta_{CP_m+1})$ est un vecteur aléatoire de phase-types multivariées avec représentation (α_β, A). De plus, nous avons $0 < T^\beta_{\beta+1} < T^\beta_{\beta+2} < \ldots < T^\beta_{CP_m+1}$.*

Preuve Nous reprenons la définition A.2.2 donnée dans l'annexe A.2 :

1. D'après la définition des ensembles Γ_k, il est clair que $\bigcap_{k=\beta+1}^{CP_m+1} \Gamma_k = \Delta$.
2. selon Neuts (1975), l'absorption dans Δ est certaine presque sûrement (= p.s.) si et seulement si A n'est pas singulière. C'est le cas ici puisque $\det A = (-1)^{CP_m} \prod_{k=1}^{CP_m} \lambda_k \neq 0$.
3. étant donné qu'à l'instant $t = 0$ le processus se situe dans l'état β, il en découle que $T_k^\beta > 0$ pour $k \in \{\beta + 1, \cdots, CP_m + 1\}$ ce qui achève la première partie de la proposition.

La deuxième partie de la proposition est tirée immédiatement du fait que $\Gamma_{CP_m+1} \subset \Gamma_{CP_m} \subset \ldots \subset \Gamma_1$.

□

L'ensemble des probabilités et paramètres intervenant dans le modèle de différenciation stochastique peut être regroupé sous la forme d'un vecteur Θ_{dif} défini de la façon suivante :

$$\Theta_{dif} = \{\lambda_i\}_{i \in \{1,...,CP_m\}} \cup \{q_{i,j}\}_{(i,j) \in \{1,...,CP_m,\Delta\}^2}$$

avec $i < j$. Dans le chapitre 5, nous proposons une méthode permettant d'estimer Θ_{dif} à partir de données expérimentales sur une population de vraies plantes.

3.3 Processus de branchement associés au modèle de développement stochastique

Dans le cas du modèle de développement stochastique $\mathcal{S}$, les variables topologiques N_n^a, $a \in \mathcal{B} \cup \mathcal{O}$ et $n \in \mathbb{N}$, sont des variables aléatoires dont l'étude stochastique est

primordiale pour la compréhension, l'identification et la validation du modèle de croissance de plante. Les techniques utilisées reposent souvent sur la décomposition de la plante en structures élementaires. Dans cette optique, nous complètons les notations déjà existantes en ajoutant des crochets qui permettent de préciser l'origine des structures étudiées. Pour $a' \in \mathcal{B} \cup \mathcal{O}$, soit alors $N_n^a[a']$ le nombre d'organes a dans une structure d'âge n initiée par un organe a' (il s'agit du nombre de bourgeons actifs de type a si $a \in \mathcal{B}$ ou alors du nombre total d'organes de type a si $a \in \mathcal{O}$). Par convention, si les crochets sont omis, alors la structure considérée a débuté par un bourgeon de CP 1 (c'est-à-dire $N_n^a = N_n^a[b_1]$). C'est le cas si cette structure est la plante toute entière. Nous en déduisons immédiatement un lemme préliminaire qui nous aidera pour la suite de cette section :

Lemme 3.3.1 *Si $o \in \mathcal{O}$, alors $N_n^a[o] = \delta_o(a)$ pour tout $a \in \mathcal{B} \cup \mathcal{O}$ et $n \geq 0$.*

Preuve Ce résultat est simplement dû au fait qu'un organe $o \in \mathcal{O}$ ne produit pas de nouveaux organes et ne change pas de nature botanique au cours de la croissance de la plante. Il n'évolue donc pas au fil des cycles de développement.

□

Le processus aléatoire sous-jacent au modèle $\mathcal{S}$ est un processus de branchement de Galton-Watson multitype ou PBGWM en abrégé (voir l'annexe A.3). Ce type de processus aléatoire a été amplement étudié par le passé (Harris (1963), Mode (1971) ou encore Athreya and Ney (2004)). En ce qui concerne le modèle $\mathcal{S}$, nous pouvons distinguer deux types de PBGWM. Le premier ne se concentre que sur la population des bourgeons actifs (= processus de branchement simple), le second prend en compte l'ensemble des organes de la plante (= processus de branchement complet). Soit alors B_n, $n \in \mathbb{N}$, le vecteur aléatoire de taille CP_m dont les composantes donnent le nombre de bourgeons actifs dans la plante au cycle de développement n :

$$B_n = (N_n^b)_{b \in \mathcal{B}}.$$

Alors, dans ce cas, nous avons :

Théorème 3.3.2 (Processus de branchement simple) *Le processus $(B_n)_{n \geq 0}$ est un processus de branchement de Galton-Watson multitype.*

Preuve Le nombre de bourgeons actifs b au cycle $n+1$, N_{n+1}^b, est le résultat de l'étape d'organogenèse du cycle n. Il s'agit donc de l'ensemble des bourgeons produits par les bourgeons de la plantes au cycle n (les organes $o \in \mathcal{O}$ ne peuvent produire de bourgeons). Sachant qu'un bourgeon b' produit $N_1^b[b']$ bourgeons b, nous en déduisons :

$$N_{n+1}^b = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} N_1^b[b']^{(k)}$$

avec $N_1^b[b']^{(1)}$, ..., $N_1^b[b']^{(N_n^{b'})}$, $N_n^{b'}$ copies indépendantes de $N_1^b[b']$. L'indépendance de ces variables aléatoires est due au fait que les organes se comportent de façon indépendante les uns des autres.

N.B. 3.2 Si $N_n^{b'} = 0$, alors la somme est nulle par convention.

Finalement, en regroupant les variables aléatoires sous la forme d'un vecteur, il vient :

$$B_{n+1} = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} (N_1^b[b'])_{b \in \mathcal{B}}^{(k)} = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} B_1[b']^{(k)}$$

avec $B_1[b'] \stackrel{\text{def}}{=} (N_1^b[b'])_{b \in \mathcal{B}}$. Les vecteurs aléatoires $B_1[b']^{(1)}$, ..., $B_1[b']^{(N_n^{b'})}$ sont bien $N_n^{b'}$ copies indépendantes d'un même vecteur aléatoire toujours à cause de l'indépendance de comportement des bourgeons. Ainsi, d'après la définition A.3.1, le processus $(B_n)_{n \geq 0}$ est un PBGWM.

□

Le processus de branchement $(B_n)_{n \geq 0}$ est dit simple puisqu'il ne prend en compte que les bourgeons actifs (les types des individus correspondent à l'ensemble des symboles de $\mathcal{B}$). Il s'agit du processus à étudier lorsque le centre d'intérêt est la dynamique de croissance de la plante. Nous pouvons mettre en évidence un deuxième processus de branchement $(N_n)_{n \geq 0}$, dit complet, qui se concentre sur la dynamique d'évolution de la plante toute entière :

Théorème 3.3.3 (Processus de branchement complet) *Le processus $(N_n)_{n \geq 0}$ est un processus de branchement de Galton-Watson multitype.*

Preuve Le vecteur N_{n+1} comprend les bourgeons actifs au cycle $n+1$ ($= B_{n+1}$) ainsi que les organes constitutifs de la plante après l'étape d'organogenèse du cycle n (c'est-à-dire $(N_{n+1}^o)_{o \in \mathcal{O}}$). En ce qui concerne les premiers, nous avons déjà vu dans la démonstration du théorème 3.3.2 que :

$$N_{n+1}^b = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} N_1^b[b']^{(k)}.$$

Comme $N_1^b[o] = 0$ si $o \in \mathcal{O}$ (cf le lemme 3.3.1), nous avons alors :

$$N_{n+1}^b = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} N_1^b[b']^{(k)} + \sum_{o \in \mathcal{O}} \sum_{k=1}^{N_n^o} N_1^b[o]^{(k)} = \sum_{a \in \mathcal{B} \cup \mathcal{O}} \sum_{k=1}^{N_n^a} N_1^b[a]^{(k)} \tag{3.1}$$

avec $N_1^b[a]^{(1)}$, ..., $N_1^b[a]^{(N_n^a)}$, N_n^a copies indépendantes de $N_1^b[a]$. Considérons maintenant le nombre d'organes $o \in \mathcal{O}$ de la plante au début du cycle $n+1$: N_{n+1}^o. Il s'agit en fait de la somme des organes $o \in \mathcal{O}$ de la plante au cycle n ($= N_n^o$) et de ceux créés lors de l'étape d'organogenèse du même cycle :

$$N_{n+1}^o = N_n^o + \text{nouveaux organes } o \text{ du cycle } n.$$

Les nouveaux organes du cycle n sont ceux créés par les bourgeons actifs lors de l'organogenèse du même cycle :

$$\text{nouveaux organes } o \text{ du cycle } n \quad = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} N_1^o[b']^{(k)}$$

avec $N_1^o[b']^{(1)}$, ..., $N_1^o[b']^{(N_n^{b'})}$, $N_n^{b'}$ copies indépendantes de $N_1^o[b']$. L'indépendance est toujours due à l'indépendance de comportement des bourgeons lors de l'organogenèse. Ensuite, en utilisant à nouveau le lemme 3.3.1, nous avons :

$$N_n^o = \sum_{k=1}^{N_n^o} N_1^o[o] = \sum_{o' \in \mathcal{O}} \sum_{k=1}^{N_n^{o'}} N_1^o[o']^{(k)}$$

ce qui donne :

$$N_{n+1}^o = \sum_{b' \in \mathcal{B}} \sum_{k=1}^{N_n^{b'}} N_1^o[b']^{(k)} + \sum_{o' \in \mathcal{O}} \sum_{k=1}^{N_n^{o'}} N_1^o[o']^{(k)} = \sum_{a \in \mathcal{B} \cup \mathcal{O}} \sum_{k=1}^{N_n^a} N_1^o[a]^{(k)}. \tag{3.2}$$

Finalement, en regroupant les équations (3.1) et (3.2) sous la forme d'un vecteur avec $a \in \mathcal{B} \cup \mathcal{O}$, nous avons :

$$N_{n+1} = \sum_{a \in \mathcal{B} \cup \mathcal{O}} \sum_{k=1}^{N_n^a} (N_1^{a'}[a])_{a' \in \mathcal{B} \cup \mathcal{O}}^{(k)} = \sum_{a \in \mathcal{B} \cup \mathcal{O}} \sum_{k=1}^{N_n^a} N_1[a]^{(k)}$$

avec $N_n[a] \stackrel{\text{def}}{=} (N_n^{a'}[a])_{a' \in \mathcal{B} \cup \mathcal{O}}$ pour $a \in \mathcal{B} \cup \mathcal{O}$. Pour tout a, les vecteurs $N_1[a]^{(1)}$,..., $N_1[a]^{(N_n^a)}$ sont N_n^a copies indépendantes de $N_1[a]$ (indépendance du comportement). Ainsi, d'après la définition A.3.1, le processus $(N_n)_{n \geq 0}$ est un PBGWM.

□

L'étude de $(N_n)_{n \geq 0}$ permet entre autre de calculer les distributions et plus particulièrement les moments associés au nombre d'organes $o \in \mathcal{O}$ d'un type donné (entrenœuds, feuilles, fruits, ...) grâce à l'utilisation de fonctions génératrices, voir la section 3.4. Il est possible également d'exprimer la distribution des descendants à l'aide des probabilités introduites dans la section 3.1 (ceci est entre autre effectué dans les démonstrations de la section 3.4).

3.4 Fonctions génératrices

Les fonctions génératrices (cf l'annexe A.1) sont un outil d'étude probabiliste très efficace permettant entre autre de calculer les moments à tout ordre d'une variable aléatoire. Dans cette section, nous déterminons la fonction génératrice associée au modèle de développement stochastique $\mathcal{S}$ (voir Loi and Cournède (2008) et Loi and Cournède (2008)). Pour ce faire, nous déterminons tout d'abord la fonction génératrice associée à l'organogenèse puis celle associée à la différenciation. Ensuite, nous calculons récursivement la fonction génératrice de $\mathcal{S}$ grâce à la composition des fonctions génératrices précédentes. Dans un soucis de clarté des expressions mathématiques, nous supposerons que $\mathcal{O}$ ne contient que les symboles liés aux entrenœuds (ce qui n'est pas gênant puisque, pour la plupart des plantes, le nombre d'organes d'un autre type est proportionnel au nombre d'entrenœuds) :

$$\mathcal{O} = \{e_i\}_{i \in \{1,...,CP_m\}}.$$

S'il y a besoin de préciser la classe physiologique j des axes portés par un entrenœud de CP i, un deuxième indice j sera ajouté au symbole e_i le décrivant. Cet entrenœud sera alors désigné par $e_{i,j}$. Par soucis de simplicité et de clarté d'écriture, les symboles utilisés pour désigner les éléments de la plante seront aussi utilisés comme variables des fonctions génératrices. Soient alors $\mathbf{b} = (b_1, \ldots, b_{CP_m})$ et $\mathbf{e} = (e_1, \ldots, e_{CP_m})$ deux vecteurs de $\mathbb{N}^{CP_m}$. Nous avons donc :

Théorème 3.4.1 (Fonction génératrice de l'organogenèse) *La fonction génératrice de l'organogenèse (sans différenciation) $\Psi_1^{org}[a]$ associée à $N_1[a]$, $a \in \mathcal{B} \cup \mathcal{O}$, est donnée par :*

– *si $a = b_i \in \mathcal{B}$:*

$$\begin{aligned}\Psi_1^{org}[b_i](\mathbf{b},\mathbf{e}) =& 1 - p_s(i) + p_s(i)(1-p_a(i))b_i \\ &+ p_s(i)p_a(i)\left[\sum_{m\in\mathbb{N}} \sum_{m_0+m_i+\ldots+m_{CP_m}=m} P_{e_i}(m_0, \mathbf{0}_{i-1}, m_i, \ldots, m_{CP_m}) e_i^m b_i \right. \\ &\left. \times \prod_{j=i}^{CP_m} \left\{ \sum_{l_j\in\mathbb{N}} \mathbb{P}(N_1^{b_j}[b_i] = l_j | N_1^{e_{i,j}}[b_i] = m_j) b_j^{l_j} \right\} \right]\end{aligned}$$

avec :

$$\mathbb{P}(N_1^{b_j}[b_i] = l_j | N_1^{e_{i,j}}[b_i] = m_j) = \begin{cases} \displaystyle\sum_{l_j^1+\ldots+l_j^{m_j}=l_j} \prod_{v=1}^{m_j} p_{b_j}^{e_i}(l_j^v) & si \;\; m_j \geq 1 \\ 1 & si \;\; m_j = l_j = 0 \\ 0 & sinon. \end{cases}$$

– *si $a = e_i \in \mathcal{O}$:*

$$\Psi_1^{org}[e_i](\mathbf{b},\mathbf{e}) = e_i.$$

Preuve Considérons un modèle $\mathcal{S}$ avec organogenèse et sans différenciation. La fonction génératrice associée à $N_1[a]$, avec $a \in \mathcal{B} \cup \mathcal{O}$, vaut :

$$\Psi_1^{org}[a](\mathbf{b},\mathbf{e}) = \sum_{\substack{(l_1,\ldots,l_{CP_m})\in\mathbb{N}^{CP_m}, \\ (m_1,\ldots,m_{CP_m})\in\mathbb{N}^{CP_m}}} \mathbb{P}\Big(N_1[a] = (l_1,\ldots,l_{CP_m}, m_1,\ldots,m_{CP_m})\Big) \prod_{j=1}^{CP_m} b_j^{l_j} \prod_{j=1}^{CP_m} e_j^{m_j}.$$

Pour prouver le théorème, il suffit de déterminer la distribution de $N_1[a]$. Le cas le plus simple correspond à $a = e_i \in \mathcal{O}$. Etant donné qu'un entrenœud e_i ne produit aucun autre organe, celui-ci reste inchangé après organogenèse. On a donc $N_1[e_i] = (\mathbf{0}_{CP_m+i-1}, 1, \mathbf{0}_{CP_m-i})$ avec $\mathbf{0}_k$ le vecteur de tailles k dont toutes les composantes sont nulles. D'où :

$$\Psi_1^{org}[e_i](\mathbf{b},\mathbf{e}) = e_i$$

Si $a = b_i \in \mathcal{B}$, il faut énumérer tous les cas possibles en fonction des situations botaniquement réalisables. Soient S_i et A_i deux évènements tels que :

$$S_i = \{ \text{ le bourgeon } b_i \text{ survie } \} \qquad A_i = \{ \text{ le bourgeon } b_i \text{ est actif } \}$$

Quatre situations sont alors possibles :

1. le bourgeon b_i meurt avec une probabilité $1-\mathbb{P}(S_i) = 1-p_s(i)$. Cela se traduit par $N_1[b_i] = \mathbf{0}_{2CP_m}$.
2. le bourgeon b_i survit mais ne produit rien avec une probabilité :

$$\mathbb{P}(\bar{A}_i \cap S_i) = \mathbb{P}(\bar{A}_i|S_i)\mathbb{P}(S_i) = (1-\mathbb{P}(A_i|S_i))\mathbb{P}(S_i) = (1-p_a(i))p_s(i).$$

   Dans ce cas, on a $N_1[b_i] = (\mathbf{0}_{i-1}, 1, \mathbf{0}_{CP_m+i-1})$.
3. le bourgeon b_i survit et est actif. Il produit alors non seulement un bourgeon apical b_i mais aussi un nombre $m > 0$ d'entrenœuds e_i et un nombre l_j de bourgeons axillaires b_j avec $j \in \{i, \ldots, CP_m\}$, c'est-à-dire $N_1[b_i] = (\mathbf{0}_{i-1}, l_i + 1, \ldots, l_{CP_m}, \mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i}) \stackrel{\text{def}}{=} D$. La probabilité associée à cet évènement vaut :

$$\begin{aligned}\mathbb{P}\Big(N_1[b_i] = D\Big) &= \mathbb{P}\Big(\{N_1[b_i] = D\} \cap S_i \cap A_i\Big) = \mathbb{P}\Big(N_1[b_i] = D|S_i \cap A_i\Big)\mathbb{P}(A_i|S_i)\mathbb{P}(S_i) \\ &= \mathbb{P}\Big(N_1^{\mathbf{b}}[b_i] = (\mathbf{0}_{i-1}, l_i, \ldots, l_{CP_m})|N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})\Big)\mathbb{P}\Big(N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})\Big)p_s(i)p_a(i)\end{aligned}$$

avec $N_1^{\mathbf{b}}[b_i] = (N_1^{b_1}[b_i], \ldots, N_1^{b_{CP_m}}[b_i])$ et $N_1^{\mathbf{e}}[b_i] = (N_1^{e_1}[b_i], \ldots, N_1^{e_{CP_m}}[b_i])$. Etant donné que $m > 0$, l'évènement $\{N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})\}$ est inclue dans $S_i \cap A_i$ ce qui explique pourquoi on ne conditionne plus par rapport à $S_i \cap A_i$ dans $\mathbb{P}\Big(N_1^{\mathbf{b}}[b_i] = (\mathbf{0}_{i-1}, l_i, \ldots, l_{CP_m})|N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})\Big)$ ni dans $\mathbb{P}\Big(N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})\Big)$. Il faut maintenant différencier les entrenœuds e_i créés selon la classe physiologique j des bourgeons axillaires qu'ils portent. Nous avons alors $m = m_0 + \sum_{j=i}^{CP_m} m_j$ avec m_j le nombre d'entrenœuds de type $e_{i,j}$ si $j \geq i$ et m_0 le nombre d'entrenœuds de CP i ne portant pas de bourgeon axillaire. Ainsi :

$$\mathbb{P}\Big(N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})|S_i \cap A_i\Big) = \sum_{m_0+m_i+\ldots+m_{CP_m}=m} p_{e_i}(m_0, \mathbf{0}_{i-1}, m_i, \ldots, m_{CP_m}).$$

De plus, comme les variables aléatoires comptant le nombre de bourgeons axillaires portés par un entrenœud sont indépendantes, nous en obtenons également :

$$\mathbb{P}\Big(N_1^{\mathbf{b}}[b_i] = (\mathbf{0}_{i-1}, l_i, \ldots, l_{CP_m})|N_1^{\mathbf{e}}[b_i] = (\mathbf{0}_{i-1}, m, \mathbf{0}_{CP_m-i})\Big) = \prod_{j=i}^{CP_m} \mathbb{P}(N_1^{b_j}[b_i] = l_j|N_1^{e_{i,j}}[b_i] = m_j).$$

Ainsi, nous en déduisons :

$$\begin{aligned}\mathbb{P}\Big(N_1[b_i] = D\Big) = p_s(i)p_a(i)\Big[&\sum_{m_0+m_i+\ldots+m_{CP_m}=m} p_{e_i}(m_0,\mathbf{0}_{i-1}, m_i, \ldots, m_{CP_m}) \\ &\times \prod_{j=i}^{CP_m} \mathbb{P}(N_1^{b_j}[b_i] = l_j|N_1^{e_{i,j}}[b_i] = m_j)\Big].\end{aligned}$$

Si $m_j = 0$, alors le bourgeon b_i ne produit aucun bourgeon b_j (puisqu'il ne produit aucun entrenœud de type $e_{i,j}$). Dans ce cas, $\mathbb{P}(N_1^{b_j}[b_i] = l_j | N_1^{e_{i,j}}[b_i] = m_j) = 1$ si $l_j = 0$ et vaut 0 dans tous les autres cas. Si $m_j \geq 1$, alors le bourgeon b_i produit l_j bourgeons b_j qui seront portés par m_j entrenœuds de type $e_{i,j}$. Notons l_j^v le nombre de bourgeons b_j portés par le v-ième entrenœud $e_{i,j}$. Nous avons donc $l_j = \sum_{j=1}^{CP_m} l_j^v$. Ainsi, étant donné que les nombres de bourgeons axillaires portés par un entrenœud sont indépendants, nous obtenons :

$$\mathbb{P}(N_1^{b_j}[b_i] = l_j | N_1^{e_{i,j}}[b_i] = m_j) = \sum_{l_j^1 + \ldots + l_j^{m_j} = l_j} \prod_{v=1}^{k_j} p_{b_j}^{e_i}(l_j^v) \qquad \text{si} \quad m_j \geq 1.$$

4. le bourgeon produit une situation botanique impossible (par exemple, un bourgeon b_i qui produirait des entrenœud e_j avec $j \neq i$ ou qui produit des bourgeons b_j avec $j < i$). La probabilité associée à cette situation est nulle.

Nous en déduisons la fonction génératrice liée à l'organogenèse $\Psi_1^{org}[b_i]$ en sommant sur toutes les valeurs de $N_1[b_i]$ possibles (qui sont décrites par les situations précédentes) et en permutant les signes somme et produit (ce qui est possible puisque le nombre de termes est fini).

□

N.B. 3.3 Si l'on ne s'intéresse qu'à la dynamique de croissance (c'est-à-dire au processus de branchement simple $(B_n)_{n\geq 0}$), la fonction génératrice associée à $B_1[a]$ s'obtient simplement en attribuant au vecteur $\mathbf{e}$ de $\Psi_1^{org}[a](\mathbf{b}, \mathbf{e})$ la valeur $\mathbf{1}_{CP_m}$ avec $\mathbf{1}_{CP_m}$ le vecteur de taille CP_m toutes les composantes valent 1. Afin de simplifier les notations, cette fonction génératrice sera encore désignée par $\Psi_1^{org}[a](\mathbf{b})$.

Grâce aux phase-types introduites dans la section 3.2, nous pouvons écrire la fonction génératrice associée à la différenciation. Nous rappelons que le vecteur des temps d'atteinte des classes physiologiques est un vecteur aléatoire de phase-types multivariées avec représentation (α_β, A) (cf la proposition 3.2.1). Soit $b_{\beta\to\phi}^k$ le symbole (et la variable) associé à un bourgeon de CP ϕ qui est apical pour un axe d'âge k qui a débuté par un bourgeon de CP β (voir la figure 3.4). Etant donné que les symboles $b_{\beta\to\phi}^k$ dépendent de l'âge k des axes, nous travaillerons avec un intervalle de temps fini. Ainsi, nous introduisons l'âge de la plante notée t_m. L'indice k varie donc entre 0 et t_m. Cette notation n'est conservée que dans la présente section.

Notons $\mathcal{B}^{dif} = \{b_{\beta\to\phi}^k\}_{(\beta,\phi)\in\{1,\ldots,CP_m\}^2, k\in\{0,\ldots,t_m\}}$ l'ensemble des symboles associés aux bourgeons pour la différenciation et $\mathbf{b}^{dif} = (b_{\beta\to\phi}^k)_{(\beta,\phi)\in\{1,\ldots,CP_m\}^2, k\in\{0,\ldots,t_m\}}$ le vecteur de variables associé. Soit D l'évènement $\{T_\phi^\beta < k-1, T_{\phi+1}^\beta > k-1\}$. Soit g_k la matrice diagonale de taille $CP_m \times CP_m$ dont le i-ème élément diagonal vaut 1 si $i \in \Gamma_k^c$ et 0 sinon avec Γ_k^c le complémentaire de Γ_k dans $\mathcal{E}$. Nous avons alors :

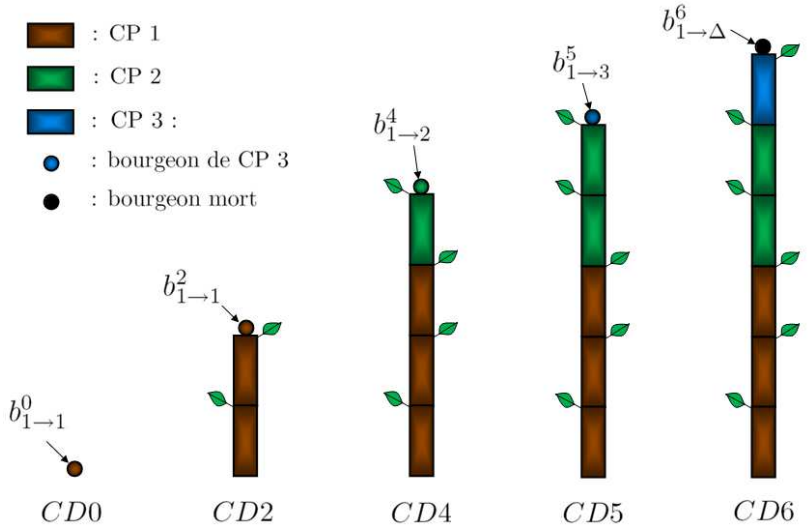

FIG. 3.4 – Notations pour la différenciation. Δ correspond à l'état final de différenciation (ici la mort du bourgeon apical). CP = Classe Physiologique. CD = Cycle de Développement.

Théorème 3.4.2 (Fonction génératrice de la différenciation) *La fonction génératrice associée à la différenciation (sans organogenèse)* $\Psi_1^{dif}[a](\mathbf{b}^{dif}, \mathbf{e})$*, avec* $a \in \mathcal{B}^{dif} \cup \mathcal{O}$*, est donnée par :*

– *si* $a = b^k_{\beta\to\phi} \in \mathcal{B}^{dif}$ *:*

$$\Psi_1^{dif}[b^k_{\beta\to\phi}](\mathbf{b}^{dif}, \mathbf{e}) = \mathbb{P}(T^\beta_{CP_m+1} < k|D) + \mathbb{P}(T^\beta_{\phi+1} > k|D) b^k_{\beta\to\phi} + \sum_{i=1}^{CP_m-\phi} \mathbb{P}(T^\beta_{\phi+i} < k, T^\beta_{\phi+i+1} > k|D) b^k_{\beta\to\phi+i}$$

si $1 \le k \le t_m$ *et par :*

$$\Psi_1^{dif}[b^0_{\beta\to\beta}](\mathbf{b}^{dif}, \mathbf{e}) = b^0_{\beta\to\beta}$$

si $k = 0$.

– *si* $a = e_i \in \mathcal{O}$ *:*

$$\Psi_1^{dif}[e_i](\mathbf{b}^{dif}, \mathbf{e}) = e_i.$$

Toutes les probabilités sont des fonctions explicites de (α_β, A) *:*

$$\mathbb{P}(T^\beta_{\phi+1} > k|D) = \frac{{}^t(\alpha^\beta)[g_{\beta+1}e^{A(k-1)} - (g_{\beta+1})^{\delta_{\beta+1}(\phi)}e^{A(k-1)}g_\phi]e^A g_{\phi+1}\mathbf{1}_{CP_m}}{{}^t(\alpha^\beta)[g_{\beta+1}e^{A(k-1)} - (g_{\beta+1})^{\delta_{\beta+1}(\phi)}e^{A(k-1)}g_\phi]g_{\phi+1}\mathbf{1}_{CP_m}}$$

$$\mathbb{P}(T^\beta_{\phi+i} < k, T^\beta_{\phi+i+1} > k|D) = \frac{{}^t(\alpha^\beta)[g_{\beta+1}e^{A(k-1)}g_{\phi+1} - (g_{\beta+1})^{\delta_{\beta+1}(\phi)}e^{A(k-1)}g_\phi]e^A[g_{\phi+i+1} - g_{\phi+i}]\mathbf{1}_{CP_m}}{{}^t(\alpha^\beta)[g_{\beta+1}e^{A(k-1)} - (g_{\beta+1})^{\delta_{\beta+1}(\phi)}e^{A(k-1)}g_\phi]g_{\phi+1}\mathbf{1}_{CP_m}}$$

avec $\mathbf{1}_{CP_m}$ *le vecteur de taille* CP_m *toutes les composantes valent 1.*

Preuve Soit $N_1^{b,dif}[a]$, avec a et b dans $\mathcal{B}^{dif} \cup \mathcal{O}$, la variable aléatoire donnant le nombre d'organes b résultant de la différenciation de a (l'organogenèse n'est pas prise en compte) et notons

$$N_1^{dif}[a] = \left((N_1^{b^k_{i\to j},dif}[a])_{i,j,k}\,,\, (N_1^{e_j,dif}[a])_j \right).$$

La fonction génératrice associée à la différenciation (c'est-à-dire associée à $N_1^{dif}[a]$) s'écrit alors :

$$\Psi_1^{dif}[a](\mathbf{b}^{dif},\mathbf{e}) = \sum_{\left((l^k_{i\to j})_{i,j,k}\,,\,(m_j)_j\right)} \mathbb{P}\Big(N_1^{dif}[a] = \Big((l^k_{i\to j})_{i,j,k}\,,\,(m_j)_j\Big)\Big) \prod_{i=1}^{CP_m}\prod_{j=1}^{CP_m}\prod_{k=0}^{t_m}(b^k_{i\to j})^{l^k_{i\to j}} \prod_{j=1}^{CP_m} e_j^{m_j}.$$

Dans la fonction génératrice précédente, la variable $l^k_{i\to j}$ est associée au nombre de bourgeons de type $b^k_{i\to j}$.

Si $a = e_i \in \mathcal{O}$, alors le résultat de la différenciation reste e_i (un entrenœud ne se différencie pas) d'où $N_1^{a,dif}[e_i] = \delta_{e_i}(a)$ p.s.. La fonction génératrice associée vaut donc :

$$\Psi_1^{dif}[e_i](\mathbf{b}^{dif},\mathbf{e}) = e_i.$$

Si $a = b^k_{\beta\to\phi}$, nous procédons de la même façon que pour le théorème 3.4.1 en énumérant toutes les situations qui sont botaniquement possibles. L'évènement $D = \{T^\beta_\phi < k-1,\ T^\beta_{\phi+1} > k-1\}$ traduit le fait que, lorsque l'axe a atteint un âge $k-1$, le bourgeon apical a une CP ϕ. Il représente la situation actuelle du bourgeon ce qui explique pourquoi les probabilités qui interviennent dans la fonction génératrice sont conditionnées par D. Les différentes situations sont :

1. Le bourgeon $b^k_{\beta\to\phi}$ ne se différencie pas (c'est-à-dire $N_1^{a,dif}[b^k_{\beta\to\phi}] = \delta_{b^k_{\beta\to\phi}}(a)$). Cela signifie que le processus des classes physiologiques ϕ^β est dans Γ_ϕ à l'instant $k-1$ mais n'a pas encore atteint $\Gamma_{\phi+1}$ à l'instant k (sinon la CP du bourgeon apical serait supérieure ou égale à $\phi+1$). La probabilité d'occurrence vaut $\mathbb{P}(T^\beta_{\phi+1} > k|D)$.

2. Le bourgeon $b^k_{\beta\to\phi}$ se différencie et atteint l'état final de différenciation (sa mort ou sa transformation en fleur, c'est-à-dire $N_1^{a,dif}[b^k_{\beta\to\phi}] = 0$ pour tout $a \in \mathcal{B}^{dif}\cup\mathcal{O}$). Dans ce cas, ϕ^β atteint Γ_{CP_m+1} avant l'instant k. La probabilité d'occurrence vaut $\mathbb{P}(T^\beta_{CP_m+1} < k|D)$.

3. Le bourgeon $b^k_{\beta\to\phi}$ se différencie et sa classe physiologique passe de ϕ à $\phi+i$ (c'est-à-dire $N_1^{a,dif}[b^k_{\beta\to\phi}] = \delta_{b^k_{\beta\to\phi+i}}(a)$). ϕ^β a donc atteint $\Gamma_{\phi+i}$ avant l'instant k mais pas $\Gamma_{\phi+i+1}$. La probabilité d'occurrence vaut alors $\mathbb{P}(T^\beta_{\phi+i} < k, T^\beta_{\phi+i+1} > k|D)$.

4. Toutes les autres situations ne sont pas possibles et donc ont une probabilité d'occurrence nulle.

En sommant toutes les situations précédentes pondérées par leur probabilité d'occurrence, nous obtenons la fonction génératrice associée à la différenciation. Le cas où $a = b^0_{\beta\to\beta}$ est particulier puisqu'il correspond à celui d'un bourgeon qui vient juste d'être créé (son âge est de 0). Dans ce cas, ce bourgeon ne peut pas se différencier immédiatement et donc $N_1^{a,dif}[b^0_{\beta\to\beta}] = \delta_{b^0_{\beta\to\beta}}(a)$.

Les probabilités écrites ci-dessus peuvent s'écrire comme des fonctions explicites de (α_β, A). En effet, d'après la section A.2, la fonction de répartition du vecteur aléatoire $(T^\beta_{\beta+1}, \dots, T^\beta_{CP_m+1})$ vaut :

$$\begin{aligned}\mathbb{P}(T^\beta_{\beta+1} > t_{\beta+1}, \dots, T^\beta_{CP_m+1} > t_{CP_m+1}) \\ = {}^t(\alpha^\beta) e^{At_{\beta+1}} g_{\beta+1} e^{A(t_{\beta+2}-t_{\beta+1})} \dots g_{CP_m} e^{A(t_{CP_m+1}-t_{CP_m})} g_{CP_m+1} \mathbf{1}_{CP_m}.\end{aligned}$$

En particulier, comme $T^\beta_{\beta+1} < T^\beta_{\beta+2} < ... < T^\beta_{CP_m+1}$ (cf la proposition 3.2.1), nous avons pour tout $\phi \in \{\beta+1, \dots, CP_m\}$, $i \in \{1, \dots, CP_m + 1 - \phi\}$ and $k \geq 0$,

$$\begin{aligned}\mathbb{P}(T^\beta_\phi > k, T^\beta_{\phi+i} > k+1) = \\ \mathbb{P}(T^\beta_{\beta+1} > 0, \cdots, T^\beta_{\phi-1} > 0, T^\beta_\phi > k, \dots, T^\beta_{\phi+i-1} > k, T^\beta_{\phi+i} > k+1, \dots, T^\beta_{CP_m} > k+1)\end{aligned}$$

et ensuite,

$$\mathbb{P}(T^\beta_\phi > k) = {}^t(\alpha^\beta)(g_{\beta+1})^{\mathbf{1}_{\{\beta+1\}}(\phi)} e^{Ak} g_\phi \mathbf{1}_{CP_m}$$

$$\mathbb{P}(T^\beta_\phi > k, T^\beta_{\phi+i} > k+1) = {}^t(\alpha^\beta)(g_{\beta+1})^{\mathbf{1}_{\{\beta+1\}}(\phi)} e^{Ak} g_\phi e^A g_{\phi+i} \mathbf{1}_{CP_m}$$

On en déduit $\mathbb{P}(T^\beta_{\phi+1} > k|D)$ par le théorème de Bayes :

$$\mathbb{P}(T^\beta_{\phi+1} > k|D) = \frac{\mathbb{P}(T^\beta_{\phi+1} > k) - \mathbb{P}(T^\beta_\phi > k-1, T^\beta_{\phi+1} > k)}{\mathbb{P}(T^\beta_{\phi+1} > k-1) - \mathbb{P}(T^\beta_\phi > k-1)}.$$

On applique à nouveau le théorème de Bayes pour $\mathbb{P}(T^\beta_{\phi+i} < k, T^\beta_{\phi+i+1} > k|D)$ et on obtient le résultat souhaité.

□

Les théorèmes précédents permettent d'obtenir la fonction génératrice associée au modèle de développement stochastique $\mathcal{S}$ de la plante (organogenèse avec différenciation). Pour cela, le théorème 3.4.1 doit être réécrit en utilisant les variables $\{b^k_{i\to j}\}_{i,j,k}$ à la place des variables $\{b_i\}_i$:

Lemme 3.4.3 *La fonction génératrice de l'organogenèse (sans différenciation) $\Psi^{org}_1[a]$ associée à $N_1[a]$, $a \in \mathcal{B}^{dif} \cup \mathcal{O}$, est donnée par :*

– *si $a = b^k_{\beta\to\phi} \in \mathcal{B}^{dif}$:*

$$\begin{aligned}\Psi^{org}_1[b^k_{\beta\to\phi}](\mathbf{b}^{dif}, \mathbf{e}) = \; & 1 - p_s(\phi) + p_s(\phi)(1-p_a(\phi)) b^{k+1}_{\beta\to\phi} \\ & + p_s(\phi) p_a(\phi) \Bigg[\sum_{m\in\mathbb{N}} \sum_{m_0+m_\phi+\dots+m_{CP_m}=m} P_{e_\phi}(m_0, \mathbf{0}_{\phi-1}, m_\phi, \dots, m_{CP_m}) e^m_\phi b^{k+1}_{\beta\to\phi} \\ & \times \prod_{j=\phi}^{CP_m} \left\{ \sum_{l_j\in\mathbb{N}} \mathbb{P}(N^{b_j}_1[b_\phi] = l_j | N^{e_{\phi,j}}_1[b_\phi] = m_j) (b^0_{j\to j})^{l_j} \right\} \Bigg]\end{aligned}$$

avec :

$$\mathbb{P}(N_1^{b_j}[b_\phi] = l_j | N_1^{e_{\phi,j}}[b_\phi] = m_j) = \begin{cases} \sum_{l_j^1+\ldots+l_j^{k_j}=l_j} \prod_{v=1}^{k_j} p_{b_j}^{e_\phi}(l_j^v) & si \;\; k_j \geq 1 \\ 1 & si \;\; k_j = l_j = 0 \\ 0 & sinon. \end{cases}$$

– *si* $a = e_i \in \mathcal{O}$ *:*

$$\Psi_1^{org}[e_i](\mathbf{b}^{dif}, \mathbf{e}) = e_i.$$

Preuve La démonstration est pratiquement identique à celle du théorème 3.4.1 puisque les probabilités impliquées dans l'organogenèse ne dépendent que de la classe physiologique actuelle du bourgeon apical. Il faut cependant changer les variables dans l'expression de la fonction génératrice. Ainsi, lorsque le bourgeon $b_{\beta\to\phi}^k$ est vivant mais inactif, celui-ci devient $b_{\beta\to\phi}^{k+1}$ (l'âge augmente d'un cycle). Si le bourgeon $b_{\beta\to\phi}^k$ est actif, il produit non seulement un bourgeon apical $b_{\beta\to\phi}^{k+1}$ mais aussi un ensemble de m_j entrenœuds $e_{\phi,j}$, avec $j \in \{\phi, \ldots, CP_m\}$, portant l_j bourgeons axillaires $b_{j\to j}^0$ ainsi que m_0 entrenœuds de CP ϕ ne portant pas de bourgeons latéraux. Ces bourgeons axillaires débutent un axe de CP j d'où $j \to j$ et l'âge qui vaut 0.

□

Soient $\mathbf{\Psi}_1^{org} = (\Psi_1^{org}[a])_{a\in\mathcal{B}^{dif}\cup\mathcal{O}}$ et $\mathbf{\Psi}_1^{dif} = (\Psi_1^{dif}[a])_{a\in\mathcal{B}^{dif}\cup\mathcal{O}}$. Nous avons alors :

Corollaire 3.4.4 (Fonction génératrice du développement) *Soit* $\Psi_n^{dev}[a]$ *la fonction génératrice associée à* $N_n[a]$*,* $a \in \mathcal{B}^{dif} \cup \mathcal{O}$*, pour le modèle* $\mathcal{S}$ *sur* $n \leq t_m$ *cycles de développement (organogenèse et différenciation). Soit* $\mathbf{\Psi}_n^{dev} = (\Psi_n^{dev}[a])_{a\in\mathcal{B}^{dif}\cup\mathcal{O}}$*.* $\mathbf{\Psi}_n^{dev}$ *est alors donné récursivement par :*

$$\mathbf{\Psi}_{n+1}^{dev} = \mathbf{\Psi}_1^{org} \circ \mathbf{\Psi}_1^{dif} \circ \mathbf{\Psi}_n^{dev}, \qquad n \in \{0, \ldots, t_m - 1\}.$$

Preuve Calculons tout d'abord la fonction génératrice $\Psi_1^{dev}[a]$ pour $a \in \mathcal{B}^{dif} \cup \mathcal{O}$. La démonstration reposent sur la composition de L-systèmes (voir le théorème 4.2.5 du chapitre 4). Soient G^{org}, G^{dif} et G^{dev} les L-systèmes stochastiques associés respectivement à l'organogenèse, la différenciation et au développement. D'après les règles de priorité définies dans la section 2.1.7, nous avons :

$$G^{dev} = G^{org} \circ G^{dif}.$$

En utilisant le théorème 4.2.5 (théorème des fonctions génératrices pour les compositions de L-systèmes) et le lemme 3.4.3, nous obtenons bien que la fonction génératrice du développement complet est la composée de celle associée à la différenciation par celle associée à l'organogenèse :

$$\mathbf{\Psi}_1^{dev} = \mathbf{\Psi}_1^{org} \circ \mathbf{\Psi}_1^{dif}.$$

Le résultat du corollaire découle simplement de la formule de composition des fonctions génératrices (Harris (1963)) :

$$\mathbf{\Psi}_{n+1}^{dev}[a] = \mathbf{\Psi}_1^{dev}[a] \circ \mathbf{\Psi}_n^{dev}[a].$$

□

Grâce au corollaire 3.4.4, il est possible de calculer les distributions mais aussi les moments associés aux nombres d'organes d'un type donné de façon récursive (grâce au corollaire A.1.2).

3.5 Finitude de la croissance

L'étude de la finitude de la croissance est primordiale afin de valider le comportement stochastique d'un modèle d'organogenèse donné. En effet, suivant les probabilités attribuées au modèle, la croissance peut être finie ou non p.s.. Il est donc impératif de vérifier que le comportement théorique de la plante est en adéquation avec ce qui est observé dans la réalité.

Etudier la finitude de la croissance d'une plante revient en fait à étudier l'extinction du processus de branchement simple $(B_n)_{n\geq 0}$. Soit E_{xt} l'évènement correspond à l'extinction de la population des bourgeons actifs :

$$E_{xt} = \{\exists n \geq 1, \quad B_n = \mathbf{0}_{CP_m}\}.$$

Soit alors $q_i = \mathbb{P}(E_{xt}|B_0 = (\mathbf{0}_{i-1}, 1, \mathbf{0}_{CP_m - i}))$ la probabilité d'extinction de la population sachant que la génération à l'instant 0 est composée d'un bourgeon b_i. La croissance de la plante est donc finie p.s. si $q_1 = 1$. Nous allons examiner en premier lieu le cas d'une plante avec organogenèse sans différenciation et voir quelles sont les conditions donnant la finitude de la croissance. Ensuite, nous montrerons que, pour une plante avec organogenèse et différenciation mais sans réitération, la croissance est finie p.s..

Considérons tout d'abord un modèle de croissance sans différenciation. Le théorème suivant donne une condition sur la finitude de la croissance :

Théorème 3.5.1 *Soit un modèle de croissance avec organogenèse mais sans différenciation. Alors les conditions suivantes sont équivalentes :*

$$\forall i \in \{1, \ldots, CP_m\}, \quad \mathbb{E}[N_1^{b_i}[b_i]] \leq 1 \quad \Leftrightarrow \quad \textit{la croissance est finie p.s.}$$

avec $\mathbb{E}[N_1^{b_i}[b_i]]$ *l'espérance de la variable aléatoire* $N_1^{b_i}[b_i]$.

Preuve On souhaite utiliser le théorème A.3.3. Soit M la matrice de taille $CP_m \times CP_m$ dont la composante (i, j) est donnée par $\mathbb{E}[N_1^{b_j}[b_i]]$. Etant donné qu'un bourgeon de CP i ne peut engendrer que des bourgeons de CP $j \geq i$, la matrice M est triangulaire supérieure. Les valeurs propres de M sont donc sur sa première diagonale et sont positives ou nulles. La plus grande valeur propre ρ est alors donnée par :

$$\rho = \max_{i \in \{1,\ldots,CP_m\}} M_{i,i} = \max_{i \in \{1,\ldots,CP_m\}} \mathbb{E}[N_1^{b_i}[b_i]].$$

Nous appliquons alors le théorème A.3.3 de l'annexe A.3 (il n'est pas utile de recourir au théorème de Perron-Frobenius car nous avons immédiatement l'existence d'une valeur propre maximale positive) et nous en déduisons :

$$\rho > 1 \quad \Leftrightarrow \quad q_1 < 1.$$

Par contraposée, nous obtenons le résultat du théorème.

□

Le théorème suivant permet d'obtenir implicitement la probabilité de finitude de la croissance. Il s'agit d'un théorème classique de la théorie des processus de branchement (voir l'annexe A.3). Soit $\mathbf{\Psi}_1^{org}$ la fonction génératrice associée à $(B_1[a])_{a\in\mathcal{B}\cup\mathcal{O}}$ (voir le *Nota Bene* 3.3). Nous avons alors :

Théorème 3.5.2 *Soit* $\mathbf{q} = (q_i)_{i\in\{1,\dots,CP_m\}}$ *avec :*

$$q_i = \mathbb{P}(E_{xt}|B_0 = (\mathbf{0}_{i-1}, 1, \mathbf{0}_{CP_m-i})).$$

Alors $\mathbf{q}$ *vérifie la relation :*

$$\mathbf{q} = \mathbf{\Psi}_1^{org}(\mathbf{q}).$$

On peut même montrer que $\mathbf{q}$ est la plus petite solution x sur $[0,1]^{CP_m}$ de l'équation $x = \mathbf{\Psi}_1^{org}(x)$ au sens de la relation d'ordre partielle canonique (voir Athreya and Ney (2004)). Cependant, il n'est pas possible d'obtenir une expression explicite de q_1 dans le cas général du modèle d'organogenèse décrit dans la section 3.1. q_1 peut tout de même être calculée dans des cas simples de croissance. Prenons le cas d'un modèle de Leeuwenberg (voir Hallé et al. (1978)) avec probabilité de mort ($CP_m = 1$ ici) :

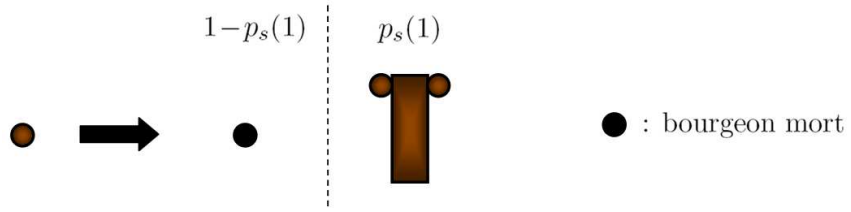

FIG. 3.5 – Organogenèse du modèle de Leeuwenberg avec probabilité de mort.

La fonction génératrice $\mathbf{\Psi}_1^{org}$ est alors donnée par :

$$\mathbf{\Psi}_1^{org}(b_1) = \Psi_1^{org}(b_1) = 1-p_s(1) + p_s(1)b_1^2$$

D'après le théorème 3.5.2, q_1 vérifie donc la relation :

$$q_1 = 1-p_s(1) + p_s(1)q_1^2$$

ce qui donne deux solutions : 1 et $p_s(1)/(1-p_s(1))$. Ainsi, si $p_s(1) < 0.5$, $q_1 = 1$. En revanche, si $p_s(1) \geq 0.5$, la croissance n'est pas finie p.s. et la probabilité d'extinction vaut alors $q_1 = p_s(1)/(1-p_s(1))$ (elle vaut par exemple 1/3 si $p_s(1) = 0.75$).

La croissance n'est donc pas toujours finie p.s. dans le cas de modèles avec organogenèse mais sans différenciation. La différenciation permet de forcer la finitude de la croissance sous certaines conditions :

Théorème 3.5.3 *La croissance est finie p.s. pour des modèles de plante avec organogenèse et différenciation ne présentant pas de réitération.*

Preuve Soit un modèle de plante sans réitération (c'est-à-dire qu'un bourgeon de CP i ne peut engendrer que des bourgeons de CP $j > i$). Introduisons les temps d'extinction des bourgeons b_i :

$$T_{ext}^{b_i} = \inf\{n \in \mathbb{N}, \quad k \geq n \quad \Rightarrow \quad N_k^{b_i} = 0\}.$$

Le temps de croissance de la plante T_{ext}^{plante} est donc le maximum des temps d'extinction des bourgeons b_i avec $i \in \{1, \ldots, CP_m\}$:

$$T_{ext}^{plante} = \max_{i \in \{1,\ldots,CP_m\}} T_{ext}^{b_i}.$$

Afin de montrer le théorème, nous allons montrer que $T_{ext}^{b_i} < \infty$ p.s. par récurrence sur i. Prenons $i = 1$. Etant donné qu'il n'y a pas de réitération, le seul bourgeon de CP 1 est la graine de la plante. Ainsi, l'extinction des bourgeons de CP 1 correspond simplement à un changement d'état botanique du bourgeon apical de l'axe principale de la plante. Comme $0 < T_2^1 < T_3^1 < \ldots < T_{CP_m+1}^1$ (cf la proposition 3.2.1), nous avons :

$$T_{ext}^{b_1} \leq T_{CP_m+1}^1.$$

Or, comme la matrice A est inversible (voir la démonstration de la proposition 3.2.1), l'absorption dans Δ (état final du bourgeon apical d'un axe : sa mort ou sa transformation en fleur) est certaine (Assaf et al. (1984)) :

$$\mathbb{P}(T_{CP_m+1}^1 < \infty) = 1$$

c'est-à-dire $T_{CP_m+1}^1 < \infty$ p.s.. On en déduit $T_{ext}^{b_1} < \infty$ p.s.. Supposons maintenant que la propriété soit vraie jusqu'au rang $i \in \{1, \ldots, CP_m - 1\}$, c'est-à-dire $T_{ext}^{b_j} < \infty$ p.s. pour tout $1 \leq j \leq i \leq CP_m - 1$. Etant donné qu'il n'y a pas de réitération, les nouveaux bourgeons axillaires de CP $i+1$ sont ceux engendrés par les bourgeons de CP $j \leq i$ (et non $j \leq i+1$ dans le cas général). Ainsi, s'il n'y a plus de bourgeons b_j avec $j \leq i$ dans la plante, aucun nouveau bourgeon de CP $i+1$ ne peut être créé. Après que le dernier bourgeon de CP $i+1$ est créé, il y a extinction des bourgeons de CP $i+1$ lorsque tout ceux-ci auront changé de nature botanique après différenciation. Comme $0 < T_{i+2}^{i+1} < \ldots < T_{CP_m+1}^{i+1}$ (cf la proposition 3.2.1), cet évenement se produit au plus tard $T_{CP_m+1}^{i+1}$ cycles après l'extinction de tous les bourgeons de CP $j \leq i$. Le temps d'extinction des bourgeons de CP $i+1$ est donc majoré par la somme du temps d'extinction de tous les bourgeons de CP $j \leq i$ et de $T_{CP_m+1}^{i+1}$:

$$T_{ext}^{b_{i+1}} \leq \max_{1 \leq j \leq i} T_{ext}^{b_i} + T_{CP_m+1}^{i+1}.$$

Comme l'absorption dans Δ est certaine, $T_{CP_m+1}^{i+1} < \infty$ p.s.. En utilisant l'hypothèse de récurrence, il vient donc que $T_{ext}^{b_{i+1}} < \infty$ p.s. ce qui prouve la récurrence. Ainsi, pour tout $i \in \{1, \ldots, CP_m\}$, $T_{ext}^{b_i} < \infty$ p.s.. Nous avons donc :

$$T_{ext}^{plante} = \max_{i \in \{1,\ldots,CP_m\}} T_{ext}^{b_i} < \infty \qquad p.s.$$

ce qui prouve le théorème.

□

Chapitre 4

Occurrences de mots dans un texte généré par un L-système stochastique

La question de l'occurrence de mots d'un type donné dans des textes générés aléatoirement a fait l'objet de nombreuses recherches durant ces vingt dernières années (voir entre autre Robin and Daudin (1999), Régnier (2000) et les référénces auxquelles ils font mention). Cette branche de la combinatoire suscite de l'engouement non seulement de par la richesse et l'intérêt des problèmes mathématiques rencontrés mais également de par les multiples applications dans des domaines tels que la télécommunication (Barbara and Imielinski (1994)), la compression de données (Szpankowski (1993), Luczak and Szpankowski (1997)) et la biologie moléculaire (avec entre autre Robin (2002) et les références auxquelles il est fait mention). Concernant ce dernier domaine, la recherche de motifs dans des séquences d'ADN (par exemple la recherche de motifs promoteurs pour mettre en évidence la position de certains gènes (Fickett (1982), Robin et al. (2002), Boeva et al. (2006), Eng et al. (2009))) a soulevé de nombreuses questions théoriques. Parmi les problèmes mathématiques rencontrés, nous pouvons citer :

- le calcul de la distribution (ou des moments) associée au nombre d'occurrences d'un mot dans un texte d'une longueur donnée (Robin and Daudin (1999), Régnier (2000)) ;
- l'étude de la répartition d'un mot le long d'un texte d'une taille donnée (le mot est-il réparti uniformément ou irrégulièrement, Karlin and Macken (1991)) ;
- la mise en évidence de certains mots d'un texte dont la fréquence d'occurrence est particulièrement basse ou élevée (Prum et al. (1995), Gelfaud (1995), Régnier and Denise (2004)).

Du point de vue modélisation mathématique, les textes apparaissent comme étant la trajectoire d'une chaîne de Markov dont l'espace d'état est l'alphabet des lettres composant les séquences d'intérêt. Par exemple, en biologie moléculaire, les séquences d'ADN sont construites lettre après lettre, chacune d'entre elles étant tirée aléatoirement dans l'alphabet des bases azotées $\{a, t, g, c\}$. Il existe plusieurs modèles de chaîne de Markov possibles (voir Régnier (2000) et Robin and Daudin (1999)) :

- modèle **M00** (Blom and Thorburn (1982)) : il s'agit du modèle le plus simple. Les lettres sont tirées indépendamment des précédentes et de façon équiprobable. Ce modèle porte aussi le nom de modèle de Bernouilli équilibré ou uniforme.
- modèle **M0** (Fudos et al. (1996)) : dans ce modèle, les lettres sont toujours tirées indépendamment des précédentes mais plus de façon équiprobable.
- modèle de Markov d'ordre 1 (**M1**, voir Chrysaphinou and Papastavridis (1990) et Régnier and Szpankowski (1998)) : les lettres ne sont plus tirées de façon indépendante. Le texte est la trajectoire d'une chaîne de Markov d'ordre 1. La probabilité d'occurrence d'une lettre dépend alors de la dernière lettre de la séquence et est donnée par les probabilités de transition de la chaîne de Markov.
- modèle de Markov d'ordre k (**Mk**) : il s'agit d'une extension du modèle précédent à l'ordre k. La probabilité d'occurrence d'une lettre donnée dépend des k précédentes lettres de la séquence.

Robin and Daudin (1999) et Régnier (2000) obtiennent des formules exactes concernant la distribution du nombre d'occurrences d'une famille de mots pour l'un des modèles probabilistes précédents grâce au calcul des fonctions génératrices correspondantes. Prenons le cas d'une famille se réduisant à un seul mot w. La méthode utilisée par Robin and Daudin (1999) est probabiliste. Ils s'intéressent d'abord à la fonction génératrice associée à la position de la première occurrence de w puis à celle associée à la distance séparant deux w successifs. A partir de ces deux fonctions génératrices, ils en déduisent celle liée à la position de la k-ième occurrence du mot w. Si la longueur du texte est fixée, alors à partir des fonctions génératrices précédentes il est possible de donner la distribution du nombre d'occurrences du mot w dans une séquence de taille donnée.

L'approche utilisée dans Régnier (2000) est combinatoire. La première étape consiste à décomposer algébriquement un texte contenant un nombre donné de mots w en utilisant des classes combinatoires dont la structure est plus simple (par exemple, l'ensemble des textes pour lesquels le mot w est le suffixe et l'ensemble des textes commençant par w et finissant par w). On obtient alors un système d'équations algébriques que l'on transforme en un système d'équations fonctionnelles faisant intervenir les fonctions génératrices associées aux classes combinatoires précédentes et à partir duquel on extrait les quantités d'intérêt.

Dans cette section, nous nous intéressons à des textes générés aléatoirement par des L-systèmes stochastiques (Lindenmayer (1968) et Prusinkiewicz and Lindenmayer (1990)). Ce sont des grammaires à réécriture parallèle dont les règles de production sont données par un ensemble de distributions de probabilité. Un L-système génère une suite de textes à partir d'un texte initial appelé axiome. A une étape de génération donnée (également appelée étape de production), le nouveau texte est obtenu en remplaçant chaque lettre du texte actuel par une autre selon les probabilités données par les règles de production. Chaque lettre évolue donc indépendamment des lettres qui l'entourent. Dans cette configuration, le texte n'est plus généré « de gauche à droite »selon un modèle de Markov comme précédemment mais il évolue intégralement à chaque étape de production. Les méthodes développées par Robin and Daudin (1999) et Régnier (2000)

ne peuvent donc plus être mises en œuvre ici. L'objectif de ce chapitre est donc de présenter une méthode combinatoire permettant de déterminer la distribution du nombre d'occurrences d'un mot donné dans un texte généré aléatoirement par des L-systèmes stochastiques après N étapes de production. La majorité des résultats présentés par la suite sont issus de Loi and Cournède (2010). Cette méthode repose sur une approche symbolique (voir les travaux de Philippe Flajolet et notamment son ouvrage de référence Flajolet and Sedgewick (2009)). La première étape consiste à écrire une bonne spécification pour les classes combinatoires d'intérêt (voir la section 4.3.1). Autrement dit, il s'agit de décomposer ces classes d'intérêt en un ensemble de classes combinatoires dont la structure est plus simple à partir de constructions admissibles. Dans cette optique, nous introduisons une nouvelle structure algébrique pour l'ensemble des classes combinatoires pondérées construites à partir d'un ensemble de mots (les textes générés par les L-systèmes stochastiques peuvent être considérés comme des mots pondérés dont le poids est la probabilité de réalisation des textes).

Dans une première section, nous présentons le cadre combinatoire associé aux ensembles de mots pondérés (Section 4.1). Nous introduisons en particulier une structure de semi-anneau se basant sur de nouveaux opérateurs union et concaténation. La section suivante (Section 4.2) présente le modèle probabiliste lié aux textes générés aléatoirement par des L-systèmes stochastiques avec entre autre le lien avec les processus de branchement multitypes. Ensuite, la section 4.3 introduit la méthode symbolique. Nous y présentons des conditions sur les opérateurs définis dans la section 4.1 pour obtenir des constructions admissibles mais aussi des théorèmes de décomposition pour les classes combinatoires d'intérêt. Ce chapite se termine par des exemples soulignant les difficultés que l'on peut rencontrer dans un tel contexte. Dans tout le chapitre, les variables (ou vecteurs) aléatoires sont définies sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.

4.1 Cadre combinatoire

4.1.1 Classes combinatoires pondérées

Nous énonçons tout d'abord quelques concepts de combinatoire que nous étendons ensuite dans un cadre plus large. Rappelons tout d'abord la notion de classe combinatoire ainsi que celle de fonction génératrice (voir Flajolet and Sedgewick (2009) pour plus de détails) :

Définition 4.1.1 (Classe combinatoire) *Une classe combinatoire (ou simplement une classe) est un ensemble fini ou dénombrable sur lequel est définie une fonction taille* $|.|$ *satisfaisant les conditions suivantes :*

- *la taille d'un élément est un entier naturel ;*
- *le nombre d'éléments d'une taille donnée est fini.*

La fonction génératrice est l'outil mathématique central de la combinatoire permettant d'étudier les classes :

Définition 4.1.2 (Fonction génératrice d'une classe combinatoire) *La fonction génératrice Ψ d'une classe C munie d'une fonction taille $|.|$ est la série formelle donnée par la forme combinatoire suivante :*

$$\Psi(z) = \sum_{t \in C} z^{|t|}.$$

En réarrangeant les termes de la fonction génératrice Ψ, nous obtenons la série entière formelle suivante :

$$\Psi(z) = \sum_{t \in C} z^{|t|} = \sum_{n \in \mathbb{N}} A_n z^n$$

avec A_n le nombre d'éléments de C ayant une taille n. Déterminer la fonction génératrice d'une classe combinatoire permet donc de compter le nombre d'éléments d'une taille donnée appartenant à cette classe . En pratique, la fonction génératrice est obtenue à partir d'une relation fonctionnelle faisant intervenir entre autre des fonctions génératrices de classes combinatoires connues (ou tout du moins plus faciles à caractériser). Cette relation est généralement établie à partir de la forme combinatoire de Ψ puisque celle-ci permet de décomposer une classe selon les caractéristiques propres à sa structure combinatoire (voir la section 4.3 à propos de la méthode symbolique).

Nous introduisons maintenant le concept de classe combinatoire pondérée. Il s'agit en fait d'une extension de la définition de classe combinatoire dans le cas où un nombre réel positif ou nul peut être associé à chacun de ses éléments :

Définition 4.1.3 (Classe combinatoire pondérée) *Soit C une classe combinatoire. Une classe combinatoire pondérée est un ensemble $WC = \{(t, p_t) \mid t \in C\}$ tel que :*

- $\forall t \in C,\ \exists!\, (t, p_t) \in C \times \mathbb{R}^+ \,/\, (t, p_t) \in WC$;
- $\sum_{t \in C} p_t < \infty$.

Pour tout $t \in C$, le nombre p_t est appelé poids associé à t.

N.B. 4.1 Si $WC = \{(t, p_t) \mid t \in C\}\}$ est une classe combinatoire pondérée, alors C est appelée classe combinatoire associée à WC.

Nous introduisons maintenant un type particulier de classe combinatoire pondérée :

Définition 4.1.4 (Classe combinatoire stochastique) *Une classe combinatoire stochastique est une classe combinatoire pondérée $SC = \{(t, p_t) \mid t \in C\}$ telle que $\sum_{t \in C} p_t = 1$.*

Le concept de fonction génératrice associée à une classe combinatoire peut facilement s'étendre au cas pondéré :

Définition 4.1.5 (Fonction génératrice d'une classe combinatoire pondérée) *Soit* $WC = \{(t, p_t) \mid t \in C\}$ *une classe combinatoire pondérée avec* C *la classe associée. Soit* $|.|$ *une fonction taille définie sur* C*. La fonction génératrice de* WC *munie de la fonction taille* $|.|$ *est la série formelle donnée par la forme combinatoire suivante :*

$$\Psi(z) = \sum_{t \in C} p_t z^{|t|}.$$

N.B. 4.2 Pour $z \in [0, 1]$, la fonction génératrice d'une classe combinatoire stochastique SC munie de la fonction taille $|.|$ coïncide avec la fonction génératrice associée à la variable aléatoire Z donnant la taille d'un élément tiré aléatoirement dans SC selon la loi multinomiale dont les paramètres sont donnés par les poids des éléments :

$$\forall z \in [0, 1], \quad \Psi(z) = \sum_{t \in C} p_t z^{|t|} = \sum_{n \in \mathbb{N}} \mathbb{P}(Z = n) z^n.$$

L'extraction des coefficients de Ψ sous forme d'une série entière permet ainsi d'obtenir la distribution de Z.

4.1.2 Ensembles de mots pondérés

Définition et fonction génératrice

Nous présentons maintenant le cadre combinatoire attaché aux ensembles de mots pondérés. Les notations suivantes sont valables pour tout le reste du chapitre. Soit $V = \{v_1, \ldots, v_m\}$ un alphabet de taille m. Soit W l'ensemble des mots construits à partir de V et W^+ l'ensemble des mots non vides ($W^+ = W \setminus \{\epsilon\}$ avec ϵ le mot vide).

Définition 4.1.6 (Fonction de comptage) *La fonction de comptage* c *est une application de* $W \times W^+$ *dans* $\mathbb{N}$ *telle que, pour tout* $(w, u) \in W \times W^+$*,* $c(w, u)$ *donne le nombre de mots* u *dans le mot* w *(les recouvrements sont pris en compte).*

Pour tout $u \in W^+$, $c(\bullet, u)$ est une fonction taille sur W (la taille d'un élément est alors donnée par le nombre de mots u qu'il contient). Ainsi, tout sous-ensemble de W est une classe combinatoire munie de cette fonction de comptage.

Définition 4.1.7 (Ensemble de mots pondérés) *Un ensemble de mots pondérés* $\mathcal{G} = \{(w, p_w) \mid w \in G\}$ *est une classe combinatoire pondérée dont la classe associée* G *est un sous-ensemble de* W*.*

La fonction génératrice donnée par la définition 4.1.5 peut donc être utilisée pour les ensembles de mots pondérés munis d'une fonction de comptage. Nous pouvons même étendre cette notion dans le cas où l'on s'intéresse au comptage de plusieurs mots :

Définition 4.1.8 (Fonction génératrice associée à une famille de mots) *Soit* $\mathcal{G} = \{(w, p_w) \mid w \in G\}$ *un ensemble de mots pondérés et* $U = \{u_1, \ldots, u_l\}$ *une famille de mots de* W^+*. La fonction génératrice de* $\mathcal{G}$ *associée à* U *est la série formelle donnée par la forme combinatoire suivante :*

$$\Psi(z_1, \ldots, z_l) = \sum_{w \in G} p_w \prod_{i=1}^{l} z_i^{c(w, u_i)}.$$

Structure algébrique

Nous définissons une structure algébrique pour l'ensemble de tous les ensembles de mots pondérés $\mathcal{W}$ construits à partir de W :

$$\mathcal{W} = \{\mathcal{G} = \{(w, p_w) \mid w \in G\} \mid \mathcal{G} \text{ est un ensemble de mots pondérés de } W\}.$$

Dans cette optique, de nouveaux opérateurs union et concaténation sont définis sur $\mathcal{W}$.

Définition 4.1.9 (Opérateur union ' + ') *Soient G et H deux sous-ensembles de W. L'union des ensembles de mots pondérés $\mathcal{G} = \{(w, p_w) \mid w \in G\}$ et $\mathcal{H} = \{(v, q_v) \mid v \in H\}$ est définie de la façon suivante :*

$$\mathcal{G} + \mathcal{H} = \left(\bigcup_{x \in G \setminus H} \{(x, p_x)\} \right) \cup \left(\bigcup_{x \in H \setminus G} \{(x, q_x)\} \right) \cup \left(\bigcup_{x \in G \cap H} \{(x, p_x + q_x)\} \right).$$

Exemple Soit $V = \{c, d\}$ un alphabet. Soient $\mathcal{G} = \{(cd, p_1), (c, p_2)\}$ et $\mathcal{H} = \{(cd, p_3), (d, p_4)\}$ deux éléments de $\mathcal{W}$. Alors :

$$\mathcal{G} + \mathcal{H} = \{(c, p_2), (d, p_4), (cd, p_1 + p_3)\}.$$

N.B. 4.3 Notons que $+$ est différent de l'opérateur union « standard » $\cup$. En effet, supposons par exemple que $\mathcal{G} \cap \mathcal{H} \neq \{\}$. Il existe alors $(w, p) \in \mathcal{G} \cap \mathcal{H}$. Dans ce cas, (w, p) est un élément de $\mathcal{G} \cup \mathcal{H}$ mais pas nécessairement de $\mathcal{G} + \mathcal{H}$ (voir l'exemple ci-dessus avec le mot cd en prenant $p_1 = p_3 = p$: (cd, p) est présent à la fois dans $\mathcal{G}$ et $\mathcal{H}$ mais pas dans $\mathcal{G} + \mathcal{H}$).

N.B. 4.4 $+$ est une loi de composition interne pour $\mathcal{W}$. Elle est également associative, commutative et possède l'ensemble vide $\{\}$ comme élément neutre.

Par la suite, nous utiliserons la convention de notation suivante :

$$\{(w_1, p_{w_1})\} + \{(w_2, p_{w_2})\} + \ldots + \{(w_n, p_{w_n})\} = \sum_{i=1}^{n} \{(w_i, p_{w_i})\}.$$

Définition 4.1.10 (Opérateur concaténation ' . ') *Soient G et H deux sous-ensembles de W. La concaténation des ensembles de mots pondérés $\mathcal{G} = \{(w, p_w) \mid w \in G\}$ et $\mathcal{H} = \{(v, q_v) \mid v \in H\}$ est définie de la façon suivante :*

$$\mathcal{G}.\mathcal{H} = \sum_{(w,v) \in G \times H} \{(w.v, p_w q_v)\}$$

avec $w.v$ la concatenation (au sens classique) des mots w et v.

Exemple Soit $V = \{c, d\}$ un alphabet. Soient $\mathcal{G} = \{(d, p_1), (dc, p_2)\}$ et $\mathcal{H} = \{(cc, p_3), (c, p_4)\}$ deux éléments de $\mathcal{W}$. Alors :

$$\begin{aligned}\mathcal{G}.\mathcal{H} &= \{(dcc, p_1p_3)\} + \{(dc, p_1p_4)\} + \{(dccc, p_2p_3)\} + \{(dcc, p_2p_4)\} \\ &= \{(dc, p_1p_4), (dccc, p_2p_3), (dcc, p_1p_3 + p_2p_4)\}.\end{aligned}$$

Par convention, nous posons $\mathcal{F}.\{\} = \{\}.\mathcal{F} = \{\}$. Ainsi, $\mathcal{W}$ a une structure de semi-anneau (voir Duchamp et al. (2005) et Klima and Polak (2008) pour des exemples de semi-anneaux en théorie des automates et des languages).

Définition 4.1.11 (Semi-anneau) $(S, +, ., 0, 1)$ *est un semi-anneau si :*

- $(S, +, 0)$ *est un monoïde commutatif avec* 0 *comme élément neutre ;*
- $(S, ., 1)$ *est un monoïde avec* 1 *comme élément neutre ;*
- *la multiplication est distributive par rapport à l'addition ;*
- 0 *est un élément absorbant de* S *par rapport à la multiplication.*

Dans ce cas, nous avons :

Théorème 4.1.1 $(\mathcal{W}, \ + \ , \ . \ , \{\}, \{(\epsilon, 1)\})$ *est un semi-anneau. L'ensemble vide* $\{\}$ *et* $\{(\epsilon, 1)\}$ *sont respectivement les éléments neutres pour ' + ' et ' . '.*

Preuve La preuve est immédiate et repose sur des manipulations élémentaires des opérateurs ' + ' et ' . '. Notons tout de même que $w.\epsilon = \epsilon.w = w$. La quatrième propriété de la structure de semi-anneau (voir la définition 4.1.11) est due au fait que, pour tout $\mathcal{G} \in \mathcal{W}$, $\mathcal{G}.\{\} = \{\}.\mathcal{G} = \{\}$.

□

4.2 Textes générés aléatoirement par des L-systèmes stochastiques

4.2.1 0L et F0L-systèmes stochastiques

Les L-systèmes sont des grammaires à réécriture parallèle. Elles ont été introduites en 1968 par Lindenmayer (Lindenmayer (1968)) et ont été particulièrement utilisées pour modéliser le développement d'organismes cellulaires et la croissance de plantes (Lindenmayer (1968), Smith (1984), Prusinkiewicz and Lindenmayer (1990) et Françon (1990)), principalement pour des besoins de visualisation graphique (Rozenberg and Salomaa (1992)). Elles ont également fait l'objet d'études mathématiques (avec entre autre Rozenberg and Salomaa (1980a) et Rozenberg and Salomaa (1980b)). Lorsque des probabilités d'occurrence sont associées aux règles de production, les L-systèmes sont dits stochastiques :

Définition 4.2.1 (0L-système stochastique) *Un 0L-système stochastique est une grammaire* $L = \langle a, \pi \rangle$ *avec :*

- $a \in W^+$ *(= l'axiome) ;*
- π *une matrice de transition de* V *dans* W *(i.e.,* $\forall (u,v) \in V \times W$, $0 \leq \pi_{u,v} \leq 1$ *et* $\sum_{w \in W} \pi_{u,w} = 1$*) avec un nombre fini de composantes non nulles.*

N.B. 4.5 Bien que différente, la définition précédente est équivalente à celle donnée dans Prusinkiewicz and Lindenmayer (1990) : l'ensemble $P_r = \{(x,y) \in V \times W \mid \pi_{x,y} > 0\}$ définit les règles de production. Ces règles représentent toutes les évolutions possibles des lettres à travers le L-système.

N.B. 4.6 De par la nature même du 0L-système stochastique, les lettres d'un mot évoluent de façon indépendante.

Un 0L-système stochastique $L = \langle a, \pi \rangle$ génère une suite de mots $\left(w_n^L\right)_{n \in \mathbb{N}}$ de W. Cette suite est initiée par l'axiome (c'est-à-dire $w_0^L = a$). Pour $n \geq 0$, w_{n+1}^L est obtenu en remplaçant indépendamment chaque lettre x de w_n^L par une séquence de lettres y avec une probabilité $\pi_{x,y}$.

Proposition 4.2.1 *Soit* $L = \langle a, \pi \rangle$ *un 0L-système stochastique. La suite de mots* $\left(w_n^L\right)_{n \in \mathbb{N}}$ *générée par* L *est une chaîne de Markov homogène à espace d'état dénombrable* W *dont la distribution initiale est donnée par la mesure de Dirac centrée en* a *(=* δ_a*) et le noyau de transition* P *par :*

$$\forall (x,y) \in W \times W, \qquad P_{x,y} = \mathbb{P}(w_{n+1}^L = y | w_n^L = x) = \sum_{\substack{(y_1, y_2, \ldots, y_n) \in W^n, \\ y_1 . y_2 \ldots y_n = y}} \prod_{i=1}^{n} \pi_{x_i, y_i}$$

avec $x = x_1 x_2 \ldots x_n$ *et* $x_i \in V$ *pour tout* $i \in \{1, \ldots, n\}$.

Preuve La propriété de Markov découle directement du principe de construction de $\left(w_n^L\right)_{n \in \mathbb{N}}$ puisque chaque mot w_{n+1}^L est construit uniquement à partir de w_n^L. Etant donné que les lettres évoluent de façon indépendante, la probabilité que le mot $x = x_1 x_2 \ldots x_n$ évolue vers le mot $y = y_1 . y_2 \ldots y_n$ est donnée par le produit des probabilités d'évolution de chacune des lettres x_i de x vers le mot y_i, c'est-à-dire $\prod_{i=1}^{n} \pi_{x_i, y_i}$. Pour obtenir $P_{x,y} = \mathbb{P}(w_{n+1}^L = y | w_n^L = x)$, il suffit de considérer toutes les combinaisons $(y_1, y_2, \ldots, y_n) \in W^n$ de sorte que $y_1 . y_2 \ldots y_n = y$ et de sommer les probabilité d'occurrence correspondantes. La chaîne de Markov est homogène puisque les probabilités de transition données par π ne dépendent pas de l'étape n de production.

□

Définissons maintenant une classe plus générale de L-systèmes appelée F0L-systèmes stochastiques, étendant la définition standard de F0L-système (Rozenberg and Salomaa (1980a), p. 89) au cas stochastique :

Définition 4.2.2 (F0L-système stochastique) *Un F0L-système stochastique est une grammaire* $L = \langle A, \pi \rangle$ *avec :*

- A *un sous-ensemble non vide de* W^+ *(appelé ensemble des axiomes de* L*) ;*

- *pour tout $a \in A$, $L[a] = \langle a, \pi \rangle$ est un 0L-système stochastique (appelé système composant de L).*

N.B. 4.7 Tous les systèmes composants $L[a]$ avec $a \in A$ ont le même noyau de transition P. Ainsi, par définition, ce noyau sera appelé noyau de transition associé au F0L-système stochastique L.

Définition 4.2.3 (Composition de F0L-systèmes stochastiques) *Soient $L = \langle W^+, \pi \rangle$, $L^1 = \langle W^+, \pi^1 \rangle$ et $L^2 = \langle W^+, \pi^2 \rangle$ trois F0L-systèmes stochastiques et P, P^1 et P^2 les trois noyaux de transition qui leurs sont associés. L est le composé de L^1 par L^2 si $P = P^1 P^2$. Dans ce cas, on note $L = L^1 \circ L^2$.*

Une conséquence immédiate de cette définition est la propriété suivante :

Propriété 4.2.2 *Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique. Pour tout $n \geq 0$, soit L^n le F0L-système stochastique donnant les règles de production de L après n étapes de production. Alors :*

$$\forall (k, k') \in \mathbb{N}^* \times \mathbb{N}^*, \qquad L^{k+k'} = L^k \circ L^{k'} \qquad .$$

Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique. Pour $a \in W^+$, la chaîne de Markov $\left(w_n^{L[a]}\right)_{n \in \mathbb{N}}$ associée au système composant $L[a]$ peut également s'interpréter comme l'évolution d'une population dont les individus à une génération n donnée sont les lettres composant w_n^L. Ces individus sont étiquetés par la lettre de l'alphabet qu'il représente. Soit alors $Z_n^{L[a]}$ le vecteur aléatoire dont les composantes donnent le nombre de chacune des lettres de l'alphabet dans un mot généré aléatoirement par le 0L-système stochastique $L[a]$ après n étapes de production :

$$Z_n^{L[a]} = (c(w_n^{L[a]}, v))_{v \in V}, \qquad n \geq 0,$$

avec c la fonction de comptage définie dans la section 4.1.2. Alors :

Théorème 4.2.3 *Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique. Pour tout $a \in W^+$, le processus $\left(Z_n^{L[a]}\right)_{n \in \mathbb{N}}$ est un processus de branchement de Galton-Watson multitype dont la distribution des descendants est donnée par π.*

Preuve Pour tout $(v, v') \in V^2$, la variable aléatoire $c(w_1^{L[v]}, v')$ donne le nombre de lettres v' engendrées par v après une étape de production de L. Ainsi, pour $n \geq 0$ et $a \in W^+$, le nombre de lettre v' présentes dans $w_{n+1}^{L[a]}$ est égal à la somme de toutes les lettres v' créées par les lettres composant $w_n^{L[a]}$:

$$c(w_{n+1}^{L[a]}, v') = \sum_{v \in V} \sum_{k=1}^{c(w_n^{L[a]}, v)} c(w_1^{L[v]}, v')^{(k)}$$

avec $c(w_1^{L[v]}, v')^{(1)}, \ldots, c(w_1^{L[v]}, v')^{(c(w_n^{L[a]}, v))}$, $c(w_n^{L[a]}, v)$ copies indépendantes de $c(w_1^{L[v]}, v')$ (l'indépendance est due à l'indépendance d'évolution des lettres). En regroupant sous la forme d'un vecteur (avec $v' \in V$), nous avons alors :

$$Z_{n+1}^{L[a]} = \sum_{v \in V} \sum_{k=1}^{c(w_n^{L[a]}, v)} Z_1^{L[v](k)}$$

avec $Z_1^{L[v](1)}, \ldots, Z_1^{L[v](c(w_n^{L[a]}, v))}$, $c(w_n^{L[a]}, v)$ vecteurs aléatoires indépendants. D'après la définition A.3.1 de l'annexe A.3, $\left(Z_n^{L[a]}\right)_{n \in \mathbb{N}}$ est un processus de branchement de Galton-Watson multitype.

□

Soit $L = \langle a, \pi \rangle$ un 0L-système stochastique de noyau de transition P. Par la suite, W_n^L désignera l'ensemble des mots pondérés possibles générés par L après n étapes de production :

$$W_n^L = \{(w, (P^n)_{a,w}) \mid w \in W\}.$$

W_n^L est une classe combinatoire stochastique.

N.B. 4.8 Notons ici que P^n symbolise la puissance n-ième de la matrice P. $(P^n)_{a,w}$ représente donc la composante (a, w) de la matrice P^n.

4.2.2 Fonction génératrice

L'objectif du chapitre est de calculer la distribution du nombre d'occurrences d'une famille de mots dans un texte généré aléatoirement par un 0L-système stochastique. Pour ce faire, l'outil combinatoire de prédilection est la fonction génératrice donnée par la définition 4.1.8 et adaptée au cas des L-systèmes stochastiques :

Définition 4.2.4 (Fonction génératrice d'un 0L-système stochastique) *Soit $L = \langle a, \pi \rangle$ un 0L-système stochastique de noyau de transition P et $U = \{u_1, \ldots, u_l\}$ une famille de mots de W^+. La fonction génératrice de L associée à U est la fonction génératrice de W_1^L associée à U :*

$$\Psi^L(z_1, \ldots, z_l) = \sum_{w \in W} P_{a,w} \prod_{i=1}^{l} z_i^{c(w,u_i)}. \tag{4.1}$$

La définition précédente peut être adaptée pour les F0L-systèmes stochastiques :

Définition 4.2.5 (Fonction génératrice d'un F0L-système stochastique) *Soit $L = \langle A, \pi \rangle$ un F0L-système stochastique et $U = \{u_1, \ldots, u_l\}$ une famille de mots de W^+. Le vecteur de fonctions génératrices de L associé à U est défini de la façon suivante :*

$$\mathbf{\Psi}^L(z_1, \ldots, z_l) = \left(\Psi^{L[a]}(z_1, \ldots, z_l)\right)_{a \in A}, \tag{4.2}$$

avec $\Psi^{L[a]}$ la fonction génératrice du système composant $L[a]$ associée à U.

Soit L^n le F0L-système stochastique associé à la n-ième étape de production de L. Pour tout $a \in A$, déterminer les coefficients de la fonction génératrice $\Psi^{L^n[a]}$ (qui est en fait la fonction génératrice de $W_n^{L[a]}$ associée à U) écrite sous forme d'une série entière permet de résoudre le problème de la distribution d'occurrences de mots. En effet, en réarrangeant les termes de l'équation 4.1, nous obtenons :

$$\begin{aligned}\Psi^{L^n[a]}(z_1,\ldots,z_l) &= \sum_{w\in W} (P^n)_{a,w} \prod_{i=1}^{l} z_i^{c(w,u_i)} \\ &= \sum_{(k_1,\ldots,k_l)\in\mathbb{N}^l} \mathbb{P}\left(c(w_n^{L[a]},u_1)=k_1,\ldots,c(w_n^{L[a]},u_l)=k_l\right) \prod_{i=1}^{l} z_i^{k_i}.\end{aligned}$$

Les coefficients $\mathbb{P}\left(c(w_n^{L[a]},u_1)=k_1,\ldots,c(w_n^{L[a]},u_l)=k_l\right)$ forment la distribution que l'on cherche à calculer.

Dans le cas où L est le composé de deux F0L-systèmes L^1 et L^2, le vecteur de fonctions génératrices de L associé à $U = V$ s'exprime facilement en fonction de ceux correspondant à L^1 et L^2. Pour cela, nous avons tout d'abord besoin du lemme suivant :

Lemme 4.2.4 *Soit $L = \langle W^+, \pi\rangle$ un F0L-système stochastique. Pour tout $a \in W^+$, soit $\Psi^{L[a]}$ la fonction génératrice de $L[a]$ associée à V. Alors :*

$$\forall a \in W^+, \qquad \Psi^{L[a]} = \prod_{i=1}^{m} \left(\Psi^{L[v_i]}\right)^{c(a,v_i)}.$$

Preuve Soit P le noyau de transition associé à L. La fonction génératrice de $L[a]$ associée à V s'écrit :

$$\Psi^{L[a]}(z_1,\ldots,z_m) = \sum_{w\in W} P_{a,w} \prod_{i=1}^{m} z_i^{c(w,v_i)}.$$

Supposons que $a = v_1^a v_2^a \ldots v_{|a|}^a$ avec $|a|$ le nombre de lettres composant a et $v_j^a \in V$, pour tout $j \in \{1,\ldots,|a|\}$. Comme les lettres évoluent de façon indépendante,

$$P_{a,w} = \sum_{\substack{(w_1,\ldots,w_{|a|})\in W^{|a|} \\ w_1\ldots w_{|a|}=w}} \prod_{j=1}^{|a|} P_{v_j^a,w_j}.$$

Si $w = w_1 \ldots w_{|a|}$ alors $c(w,v_i) = \sum_{j=1}^{|a|} c(w_j,v_i)$. Ainsi :

$$\Psi^{L[a]}(z_1,\ldots,z_m) = \sum_{w\in W} \sum_{\substack{(w_1,\ldots,w_{|a|})\in W^{|a|} \\ w_1\ldots w_{|a|}=w}} \left[\left(\prod_{j=1}^{|a|} P_{v_j^a,w_j}\right)\left(\prod_{i=1}^{m}\prod_{j=1}^{|a|} z_i^{c(w_j,v_i)}\right)\right].$$

Les produits sur i et j comportent un nombre fini de termes. Nous pouvons donc les permuter :

$$\begin{aligned}\Psi^{L[a]}(z_1,\dots,z_m) &= \sum_{w\in W}\sum_{\substack{(w_1,\dots,w_{|a|})\in W^{|a|}\\ w_1\dots w_{|a|}=w}}\left[\left(\prod_{j=1}^{|a|}P_{v_j^a,w_j}\right)\left(\prod_{j=1}^{|a|}\prod_{i=1}^{m}z_i^{c(w_j,v_i)}\right)\right]\\ &= \sum_{w\in W}\sum_{\substack{(w_1,\dots,w_{|a|})\in W^{|a|}\\ w_1\dots w_{|a|}=w}}\left[\prod_{j=1}^{|a|}\left(P_{v_j^a,w_j}\prod_{i=1}^{m}z_i^{c(w_j,v_i)}\right)\right].\end{aligned}$$

Les deux signes $\sum$ peuvent se regrouper pour ne former plus qu'une somme sur $(w_1,\dots,w_{|a|})\in W^{|a|}$:

$$\begin{aligned}\Psi^{L[a]}(z_1,\dots,z_m) &= \sum_{(w_1,\dots,w_{|a|})\in W^{|a|}}\left[\prod_{j=1}^{|a|}\left(P_{v_j^a,w_j}\prod_{i=1}^{m}z_i^{c(w_j,v_i)}\right)\right]\\ &= \prod_{j=1}^{|a|}\left[\sum_{w_j\in W}P_{v_j^a,w_j}\prod_{i=1}^{m}z_i^{c(w_j,v_i)}\right]\\ &= \prod_{j=1}^{|a|}\Psi^{L[v_j^a]}(z_1,\dots,z_m).\end{aligned}$$

Finalement, comme a comporte $c(a,v_i)$ lettres v_i, avec $i\in\{1,\dots,m\}$, nous en déduisons :

$$\Psi^{L[a]}(z_1,\dots,z_m)=\prod_{i=1}^{m}\left(\Psi^{L[v_i]}(z_1,\dots,z_m)\right)^{c(a,v_i)}.$$

□

Grâce au lemme précédent, nous en déduisons le théorème suivant :

Théorème 4.2.5 *Soient L, L^1 et L^2 trois F0L-systèmes stochastiques dont l'ensemble des axiomes est W^+ et soient $\mathbf{\Psi}$, $\mathbf{\Psi^1}$ et $\mathbf{\Psi^2}$, les vecteurs de fonctions génératrices respectifs associés à $U=V$. Si $L=L^1\circ L^2$ alors :*

$$\forall(z_1,\dots,z_m)\in[0,1]^m,\qquad \mathbf{\Psi}(z_1,\dots,z_m)=\mathbf{\Psi^1}\circ\mathbf{\Psi^2}(z_1,\dots,z_m).$$

Preuve Soient P, P^1 et P^2 les noyaux de transition associés respectivement à L, L^1 et L^2. Par définition, nous avons donc $P=P^1P^2$. Soit $a\in W^+$. Alors :

$$\begin{aligned}\Psi^{L[a]}(z_1,\dots,z_m) &= \sum_{w\in W}P_{a,w}\prod_{i=1}^{m}z_i^{c(w,v_i)}\\ &= \sum_{w\in W}\sum_{w'\in W}P^1_{a,w'}P^2_{w',w}\prod_{i=1}^{m}z_i^{c(w,v_i)}.\end{aligned}$$

Comme $\Psi^{L[a]}$ est absolument convergente sur $[0,1]^m$, nous pouvons permuter les signes sommes :

$$\Psi^{L[a]}(z_1,\ldots,z_m) = \sum_{w'\in W} P^1_{a,w'} \sum_{w\in W} P^2_{w',w} \prod_{i=1}^{m} z_i^{c(w,v_i)} \quad = \sum_{w'\in W} P^1_{a,w'} \Psi^{L^2[w']}(z_1,\ldots,z_m).$$

Grâce au lemme 4.2.4 :

$$\Psi^{L^2[w']} = \prod_{i=1}^{m} \left(\Psi^{L^2[v_i]}\right)^{c(w',v_i)}.$$

Nous avons donc :

$$\Psi^{L[a]}(z_1,\ldots,z_m) = \sum_{w'\in W} P^1_{a,w'} \prod_{i=1}^{m} \left(\Psi^{L^2[v_i]}(z_1,\ldots,z_m)\right)^{c(w',v_i)} \quad = \Psi^{L^1[a]}(\mathbf{\Psi^2}(z_1,\ldots,z_m)).$$

Finalement, en regroupant les fonctions génératrices $\Psi^{L[a]}$ sous la forme d'un vecteur avec $a \in W^+$, nous obtenons bien :

$$\mathbf{\Psi} = \mathbf{\Psi^1} \circ \mathbf{\Psi^2}.$$

□

Corollaire 4.2.6 *Soit $\mathbf{\Psi^L_n}$ le vecteur de fonctions génératrices d'un F0L-système stochastique $L = \langle W^+, \pi\rangle$ associé à V après n étapes de production. Alors :*

$$\forall (k,k') \in \mathbb{N}^* \times \mathbb{N}^*, \qquad \mathbf{\Psi^L_{k+k'}} = \mathbf{\Psi^L_k} \circ \mathbf{\Psi^L_{k'}} = \mathbf{\Psi^L_{k'}} \circ \mathbf{\Psi^L_k}.$$

Preuve Soit L^n le F0L-système stochastique associé à la n-ième étape de production de L. D'après la propriété 4.2.2, $L^{k+k'} = L^k \circ L^{k'} = L^{k'} \circ L^k$. En utilisant le théorème 4.2.5 et en remarquant que le vecteur de fonctions génératrices de L^n associé à V est $\mathbf{\Psi^L_n}$, nous en déduisons le résultat du corollaire.

□

Grâce au corollaire précédent, nous avons en particulier :

$$\forall k \in \mathbb{N}^*, \qquad \mathbf{\Psi^L_{k+1}} = \mathbf{\Psi^L_k} \circ \mathbf{\Psi^L_1} = \mathbf{\Psi^L_1} \circ \mathbf{\Psi^L_k}.$$

Ce résultat peut également être obtenu à partir du théorème 4.2.3 soulignant le lien entre 0L-systèmes stochastiques et processus de branchement multitypes. En effet, $\mathbf{\Psi^L_n}$ est également le vecteur de fonctions génératrices associé à $\left(Z_n^{L[a]}\right)_{a\in W^+}$ où $Z_n^{L[a]}$ est le vecteur de $\mathbb{N}^m$ dont la i-ème composante donne le nombre de lettres v_i dans un mot généré aléatoirement par $L[a]$ après n étapes de production. En appliquant le théorème de composition des fonctions génératrices pour des processus de branchement multitypes (théorème A.3.1 de l'annexe A.3, Harris (1963)), nous retrouvons le résultat précédent.

4.3 Mise en place d'une méthode symbolique

L'objectif de cette section est de mettre en place une méthode symbolique permettant de calculer la distribution associée au nombre d'occurrences d'une famille de mots dans un texte généré aléatoirement par un 0L-système stochastique $L = \langle a, \pi \rangle$ après N étapes de production.

N.B. 4.9 Afin de simplifier l'écriture des équations qui suivent, nous ne considérons que des familles réduites à un seul mot $u \in W^+$ (c'est-à-dire $U = \{u\}$). L'extension à des familles de plusieurs mots ne pose pas de problème particulier.

Pour résoudre ce problème, on se propose de déterminer Ψ_N^L la fonction génératrice de L^N associée à $U = \{u\}$ avec L^N le 0L-système correspondant à la N-ième étape de production. En effet, nous avons vu dans la section 4.2.2 que cette fonction génératrice peut se mettre sous la forme d'une série entière :

$$\Psi_N^L(z) = \sum_{w \in W} (P^N)_{a,w}\, z^{c(w,u)} = \sum_{k \in \mathbb{N}} \mathbb{P}\left(c(w_N^L, u) = k\right) z^k$$

avec $\left(w_n^L\right)_{n \geq 0}$ la chaîne de Markov générée par L. La suite des coefficients $\left(\mathbb{P}\left(c(w_N^L, u) = k\right)\right)_{k \geq 0}$ correspond à la distribution qui nous intéresse. Le problème est que, la plupart du temps, Ψ_N^L ne peut pas être calculée directement à cause de la complexité des structures engendrées par L. Nous mettons en place une méthode symbolique permettant alors d'obtenir Ψ_N^L de façon récursive.

Nous commençons cette section par quelques rappels de combinatoire et nous présentons ensuite les fondements de la méthode symbolique.

4.3.1 Rappels de combinatoire

Tous les concepts mentionnés dans cette annexe sont détaillés dans Flajolet and Sedgewick (2009). Soit $\mathcal{A}$ une classe combinatoire et $|.|$ une fonction taille. Soit A la fonction génératrice de $\mathcal{A}$ munie de s :

$$A(z) = \sum_{t \in \mathcal{A}} z^{|t|} = \sum_{n \in \mathbb{N}} A_n z^n$$

avec A_n le nombre d'éléments de $\mathcal{A}$ ayant une taille n. La suite $(A_n)_{n \in \mathbb{N}}$ est appelée suite de comptage de $\mathcal{A}$ (counting sequence en anglais).

Définition 4.3.1 (Construction admissible) *Soit $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}$ une collection de classes combinatoires munies d'une fonction taille $|.|$ et $B^{(1)}, \ldots, B^{(m)}$ les fonctions génératrices correspondantes. Soient $(B_n^{(1)})_{n \in \mathbb{N}}$, ... , $(B_n^{(m)})_{n \in \mathbb{N}}$ les suites de comptage associées. Soit Φ une construction qui, pour toute collection de classes $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}$, associe une nouvelle classe*

$$\mathcal{A} = \Phi[\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(m)}].$$

La construction Φ est admissible si la suite de comptage $(A_n)_{n\in\mathbb{N}}$ de $\mathcal{A}$ ne dépend que des suites de comptage $(B_n^{(1)})_{n\in\mathbb{N}},\ldots,(B_n^{(m)})_{n\in\mathbb{N}}$ de $\mathcal{B}^{(1)},\ldots,\mathcal{B}^{(m)}$. Pour une telle construction admissible, il existe un opérateur Ψ bien défini agissant sur les fonctions génératrices correspondantes $B^{(1)}(z),\ldots,B^{(m)}(z)$:

$$A(z) = \Psi[B^{(1)}(z),\ldots,B^{(m)}(z)].$$

Définition 4.3.2 (Spécification) *Une spécification pour un r-uplet $(\mathcal{A}^{(1)},\ldots,\mathcal{A}^{(r)})$ de classes combinatoires est une collection de r équations,*

$$\begin{cases} \mathcal{A}^{(1)} &= \Phi_1(\mathcal{A}^{(1)},\ldots,\mathcal{A}^{(r)}) \\ \mathcal{A}^{(2)} &= \Phi_2(\mathcal{A}^{(1)},\ldots,\mathcal{A}^{(r)}) \\ &\cdots \\ \mathcal{A}^{(r)} &= \Phi_r(\mathcal{A}^{(1)},\ldots,\mathcal{A}^{(r)}) \end{cases} \tag{4.3}$$

où Φ_1, Φ_2, …, Φ_r sont des constructions admissibles.

Formellement, le système (4.3) est une spécification itérative si celui-ci est strictement triangulaire inférieur, c'est-à-dire que $\mathcal{A}^{(1)}$ peut être exprimée uniquement à partir de classes combinatoires de bases et, pour tout $k \in \{1,\ldots,r-1\}$, la construction de $\mathcal{A}^{(k+1)}$ ne dépend que de $\mathcal{A}^{(1)}$, …, $\mathcal{A}^{(k)}$ et de classes combinatoires de base.

4.3.2 Principe, constructions admissibles et théorèmes de décomposition

La méthode symbolique est très utilisée en analyse combinatoire, voir Flajolet and Sedgewick (2009). Elle permet de calculer des fonctions génératrices associées à des classes d'intérêt. Le principe de base consiste à transformer un système d'équations combinatoires faisant intervenir les classes d'intérêt en un système d'équations fonctionnelles impliquant les fonctions génératrices correspondantes.

Nous présentons ici cette méthode dans le cadre des textes générés aléatoirement par des 0L-systèmes stochastiques. La première étape de la méthode consiste à écrire une spécification itérative appropriée pour l'ensemble des classes combinatoires stochastiques $\{W_n^L \mid n \in \{0,\ldots,N\}\}$ sous la forme d'un système d'équations combinatoires reposant sur des constructions admissibles. Ensuite, en utilisant des règles de transformation, ces équations combinatoires sont transformées en équations fonctionnelles faisant intervenir Ψ_n^L pour $n \in \{0,\ldots,N\}$. Finalement, les coefficients de Ψ_N^L sont extraits à partir de ces équations fonctionnelles.

Le point crucial de la méthode symbolique est l'écriture de la spécification. Dans cette optique, les classes W_n^L sont décomposées algébriquement en faisant intervenir les classes W_k^L avec $k \leq n$ mais aussi des classes de structure combinatoire simple. Cette décomposition doit reposer sur des constructions admissibles. Pour ce faire, nous allons utiliser la structure de semi-anneau établie dans la section 4.1.2 pour l'ensemble $\mathcal{W}$ des ensembles de mots pondérés contruits sur W. En effet, sous certaines conditions, les opérateurs union et concaténation sont des constructions admissibles :

Définition 4.3.3 (Concaténation non génératrice) *Soit $u \in W^+$ et G et H deux sous-ensembles de W. La concaténation de G et de H est dite non génératrice par rapport à u si la condition suivante est vérifiée :*

$$\forall (w,v) \in G \times H, \qquad c(w.v,u) = c(w,u) + c(v,u).$$

Dans le cas contraire, la concaténation de G et de H est dite génératrice par rapport à u.

Par définition, la concaténation de $\mathcal{G} = \{(w, p_w) \mid w \in G\}$ et de $\mathcal{H} = \{(v, q_v) \mid v \in H\}$ est dite non génératrice par rapport à u si la concaténation de G et de H l'est. Dans ce cas :

Théorème 4.3.1 (Constructions admissibles) *Soient $\mathcal{G}$, $\mathcal{H}$ et $\mathcal{I}$ trois éléments de $\mathcal{W}$. Soit $u \in W^+$ et α,β et γ les fonctions génératrices respectives de $\mathcal{G}$, $\mathcal{H}$ et $\mathcal{I}$ associées à $U = \{u\}$. Alors, l'opérateur union ' + ' est une construction admissible avec la règle de transformation suivante :*

$$\mathcal{I} = \mathcal{G} + \mathcal{H} \qquad \Longrightarrow \qquad \gamma(z) = \alpha(z) + \beta(z).$$

Supposons de plus que la concaténation de $\mathcal{G}$ et $\mathcal{H}$ est non génératrice par rapport à u. Dans ce cas, l'opérateur concaténation ' . ' est aussi une construction admissible avec la règle de transformation suivante :

$$\mathcal{I} = \mathcal{G}.\mathcal{H} \qquad \Longrightarrow \qquad \gamma(z) = \alpha(z)\beta(z).$$

Preuve Supposons que $\mathcal{G} = \{(w, p_w) \mid w \in G\}$ et que $\mathcal{H} = \{(v, q_v) \mid v \in H\}$.
1) Si $\mathcal{I} = \mathcal{G} + \mathcal{H}$. Alors, à partir des définitions 4.1.8 et 4.1.9 :

$$\begin{aligned}\gamma(z) &= \sum_{x \in G \setminus H} p_x z^{c(x,u)} + \sum_{x \in H \setminus G} q_x z^{c(x,u)} + \sum_{x \in G \cap H} (p_x + q_x) z^{c(x,u)} \\ &= \sum_{w \in G} p_w z^{c(w,u)} + \sum_{v \in H} q_v z^{c(v,u)} = \alpha(z) + \beta(z).\end{aligned}$$

2) Si $\mathcal{I} = \mathcal{G}.\mathcal{H}$. Alors, à partir des définitions 4.1.8 et 4.1.10 :

$$\gamma(z) = \sum_{(w,v) \in G \times H} (p_w q_v) z^{c(w.v,u)}.$$

Etant donné que la concaténation de $\mathcal{G}$ et $\mathcal{H}$ est non génératrice par rapport à u, alors, pour $(w, v) \in G \times H$, $c(w.v, u) = c(w, u) + c(v, u)$. Ainsi,

$$\gamma(z) = \sum_{(w,v) \in G \times H} \left(p_w z^{c(w,u)}\right)\left(q_v z^{c(v,u)}\right) = \left(\sum_{w \in G} p_w z^{c(w,u)}\right)\left(\sum_{v \in H} q_v z^{c(v,u)}\right) = \alpha(z).\beta(z).$$

□

N.B. 4.10 Contrairement à l'opérateur union classique $\cup$, il n'y a pas besoin d'imposer que $\mathcal{G} \cap \mathcal{H} = \{\}$ pour faire de l'opérateur ' + ' une construction admissible.

Bien que la structure de semi-anneau donne un cadre algébrique pour l'écriture de la spécification, elle ne fournit pas de méthode de décomposition pour W_n^L avec $n \in \{0, \ldots, N\}$. Pour obtenir une telle décomposition, nous pouvons utiliser les théorèmes 4.3.2 et 4.3.3 suivants :

Théorème 4.3.2 (Décomposition générale) *Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique. Alors :*

$$\forall a \in W^+, \quad \forall n \in \mathbb{N}, \qquad W_{n+1}^{L[a]} = \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]}.$$

Preuve Soit P le noyau de transition associé à L. Soit $x \in W_{n+1}^{L[a]}$. Alors, il existe $s \in W$ tel que $x = (s, (P^{n+1})_{a,s})$. En utilisant le théorème de Chapman-Kolmogorov, nous avons :

$$(P^{n+1})_{a,s} = \sum_{w \in W} P_{a,w}(P^n)_{w,s} = \sum_{w \in W} \pi_{a,w}(P^n)_{w,s}. \tag{4.4}$$

Ainsi :

$$\{x\} = \{(s, \sum_{w \in W} \pi_{a,w}(P^n)_{w,s})\} = \sum_{w \in W} \{(s, \pi_{a,w}(P^n)_{w,s})\} = \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.\{(s, (P^n)_{w,s})\}.$$

Or $(s, (P^n)_{w,s}) \in W_n^{L[w]}$ pour tout $w \in W$. Nous en déduisons que, pour tout $x \in W_{n+1}^{L[a]}$, $x \in \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]}$ et donc :

$$W_{n+1}^{L[a]} \subset \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]}.$$

Réciproquement, soit $x = (s, p_s) \in \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]}$. Alors, il existe une suite $\{(r_w, (P^n)_{w,r_w})\}_{w \in W}$ de $W_n^{L[w]}$ telle que :

$$\{x\} = \{(s, p_s)\} = \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.\{(r_w, (P^n)_{w,r_w})\} = \sum_{w \in W} \{(r_w, \pi_{a,w}(P^n)_{w,r_w})\}.$$

Cette dernière équation impose $s = r_w$ pour tout $w \in W$. Ainsi, en utilisant l'équation de Chapman-Kolmogorov (4.4), nous obtenons :

$$\{x\} = \sum_{w \in W} \{(s, \pi_{a,w}(P^n)_{w,s})\} = \{(s, \sum_{w \in W} \pi_{a,w}(P^n)_{w,s})\} = \{(s, (P^{n+1})_{a,s})\}.$$

Nous en déduisons que, pour tout $x \in \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]}$, $x \in W_{n+1}^{L[a]}$ et donc :

$$\sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]} \subset W_{n+1}^{L[a]}.$$

Finalement :

$$W_{n+1}^{L[a]} = \sum_{w \in W} \{(\epsilon, \pi_{a,w})\}.W_n^{L[w]}.$$

□

Théorème 4.3.3 (Indépendance d'évolution) *Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique. Soient w_1, w_2,..., w_k, k mots de W^+. Alors :*

$$\forall n \in \mathbb{N}, \quad W_n^{L[w_1.w_2.\cdots.w_k]} = W_n^{L[w_1]}.W_n^{L[w_2]}.\cdots.W_n^{L[w_k]}.$$

Preuve Soit $k = 2$. Soient m_1 et m_2 le nombre de lettres de w_1 et de w_2 respectivement. Dans ce cas, il existe $(v_1^1, \ldots, v_{m_1}^1) \in V^{m_1}$ et $(v_1^2, \ldots, v_{m_2}^2) \in V^{m_2}$ tels que $w_1 = v_1^1.\cdots.v_{m_1}^1$ et $w_2 = v_1^2.\cdots.v_{m_2}^2$. Etant donné que les lettres d'un mot évoluent de façon indépendante, nous avons :

$$W_n^{L[w_1.w_2]} = W_n^{L[v_1^1.\cdots.v_{m_1}^1.v_1^2.\cdots.v_{m_2}^2]} = W_n^{L[v_1^1]}.\cdots.W_n^{L[v_{m_1}^1]}.W_n^{L[v_1^2]}.\cdots.W_n^{L[v_{m_2}^2]}.$$

De la même façon, pour $j \in \{1, 2\}$:

$$W_n^{L[w_j]} = W_n^{L[v_1^j.\cdots.v_{m_j}^j]} = W_n^{L[v_1^j]}.\cdots.W_n^{L[v_{m_j}^j]}.$$

Ainsi,

$$W_n^{L[w_1.w_2]} = W_n^{L[w_1]}.W_n^{L[w_2]}.$$

Le résultat du théorème se démontre par une récurrence immédiate sur $k \geq 2$.

□

Les théorèmes 4.3.2 et 4.3.3 procurent des équations combinatoires qui apparaissent comme des décompositions naturelles des classes W_n^L. Cependant, leur utilisation n'aboutit pas toujours sur des constructions admissibles (voir l'exemple 4.4.2). Dans ce cas, d'autres méthodes moins systématiques doivent être employées pour obtenir une spécification appropriée.

4.3.3 Enoncé de la méthode

Soit $L = \langle A, \pi \rangle$ un F0L-système stochastique. La méthode symbolique peut être résumée par les points suivants :

- Déterminer l'objectif : calculer la distribution associée au nombre d'occurrences d'un mot $u \in W^+$ dans un texte généré aléatoirement par le système composant $L[a]$ après N étapes de production et $a \in A$.
- Ecrire la fonction génératrice de $W_N^{L[a]}$ associée à $U = \{u\}$: $\Psi_N^{L[a]}$. Les coefficients de $\Psi_N^{L[a]}$ (écrite sous la forme d'une série entière) donnent la distribution d'intérêt.
- Ecrire une spécification itérative pour les ensembles de mots pondérés $\{W_n^{L[a]} \mid n \in \{0, \ldots, N\}\}$ en utilisant des constructions admissibles obtenues à partir des opérateurs union ' + ' et concaténation ' . '. Les théorèmes 4.3.2 et 4.3.3 permettent en général d'aboutir au résultat.
- Utiliser les règles de transformation du théorème 4.3.1 et écrire un système fermé d'équations fonctionnelles impliquant $\Psi_N^{L[a]}$ pour $n \in \{0, \ldots, N\}$.
- Résoudre directement le système précédent ou bien trouver un ensemble d'équations récursives vérifiées par les coefficients de $\Psi_N^{L[a]}$ pour $n \in \{0, \ldots, N\}$.
- En déduire les coefficients de $\Psi_N^{L[a]}$.

4.4 Exemples

La méthode symbolique est illustrée ici au travers de deux exemples tirés de la botanique (voir le chapitre 5 et plus particulièrement la section 5.2 ; d'autres exemples sont également détaillés dans cette section). Le premier exemple est le calcul de la distribution associée à un mot d'une seule lettre. Nous montrons que les théorèmes de décomposition 4.3.2 et 4.3.3 donnent directement une bonne spécification ce qui nous permet d'en déduire les équations fonctionnelles vérifiées par les fonctions génératrices d'intérêt. Le deuxième exemple traite d'un mot à deux lettres. Dans ce cas, les théorèmes de décomposition 4.3.2 et 4.3.3 ne fournissent plus directement une bonne spécification puisque les équations combinatoires obtenues ne reposent plus sur des constructions admissibles. Il faut donc les réécrire sous une forme plus convenable.

4.4.1 Exemple avec un mot d'une lettre

Soit $V = \{s, m\}$ un alphabet. Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique dont les composantes de la matrice de transition π sont toutes égales à zéro à l'exception de :

$$\pi_{s,ms} = p_1 \qquad \pi_{s,ss} = 1 - p_1 \qquad \pi_{m,mm} = p_2 \qquad \pi_{m,s} = 1 - p_2$$

avec $(p_1, p_2) \in]0, 1[^2$. Le noyau de transition correspondant est noté P. Nous souhaitons calculer la distribution associée au nombre d'occurrences du mot m dans un texte généré aléatoirement par le système composant $L[s]$ après N étapes de production. Afin de résoudre ce problème, nous voulons déterminer la fonction génératrice de $W_N^{L[s]}$ associée à m :

$$\Psi_N^{[s]}(z) = \sum_{w \in W} (P^N)_{s,w} \;\; z^{c(w,m)} = \sum_{k \in \mathbb{N}} \;\; \mathbb{P}\left(c(w_N^{L[s]}, m) = k\right) \;\; z^k$$

avec $\left(w_n^{L[s]}\right)_{n \geq 0}$ la chaîne de Markov engendrée par $L[s]$. Trouvons maintenant une spécification itérative adéquate pour $\{W_n^{L[s]} \;\; | \;\; n \in \{0, \ldots, N\}\}$. Utilisons d'abord le théorème de décomposition générale 4.3.2 :

$$\forall n \in \{0, \ldots, N-1\}, \quad W_{n+1}^{L[s]} = \{(\epsilon, 1 - p_1)\}.W_n^{L[ss]} + \{(\epsilon, p_1)\}.W_n^{L[ms]}.$$

Ensuite, en appliquant le théorème d'indépendance d'évolution 4.3.3, il vient :

$$\forall n \in \{0, \ldots, N-1\}, \qquad W_{n+1}^{L[s]} = \{(\epsilon, 1-p_1)\}.W_n^{L[s]}.W_n^{L[s]} + \{(\epsilon, p_1)\}.W_n^{L[m]}.W_n^{L[s]}. \quad (4.5)$$

De la même façon, nous obtenons une équation combinatoire pour $W_{n+1}^{L[m]}$:

$$\forall n \in \{0, \ldots, N-1\}, \qquad W_{n+1}^{L[m]} = \{(\epsilon, 1 - p_2)\}.W_n^{L[s]} + \{(\epsilon, p_2)\}.W_n^{L[m]}.W_n^{L[m]}. \quad (4.6)$$

Les équations (4.5) et (4.6) forment une spécification itérative. Notons que $W_0^{L[s]} = \{(s, 1)\}$ et $W_0^{L[m]} = \{(m, 1)\}$. Ainsi, en utilisant les règles de transformation (cf le théorème 4.3.1), nous obtenons :

$$\forall n \in \{0, \ldots, N-1\}, \qquad \begin{cases} \Psi_{n+1}^{L[s]}(z) = (1 - p_1)\left(\Psi_n^{L[s]}(z)\right)^2 + p_1 \Psi_n^{L[m]}(z)\Psi_n^{L[s]}(z) \\ \Psi_{n+1}^{L[m]}(z) = (1 - p_2)\Psi_n^{L[s]}(z) + p_2\left(\Psi_n^{L[m]}(z)\right)^2 \end{cases} \quad (4.7)$$

avec $\Psi_0^{L[s]}(z) = 1$ et $\Psi_n^{L[m]}(z) = z$. Soit $p_n^k = \mathbb{P}\left(c(W_n^{L[s]}, m) = k\right)$ (resp. $q_n^k = \mathbb{P}\left(c(W_n^{L[m]}, m) = k\right)$) la probabilité d'avoir k lettres m dans un texte généré aléatoirement par $L[s]$ (resp. $L[m]$) après n étapes de production. p_n^k (resp. q_n^k) représente le coefficient devant z^k dans le développement en série entière de $\Psi_n^{L[s]}$ (resp. $\Psi_n^{L[m]}$). Grâce au système (4.7), il est possible de déterminer récursivement la distribution $\left\{p_N^k\right\}_{k\in\mathbb{N}}$ en identifiant les coefficients des fonctions génératrices vues comme séries entières :

$$\forall n \in \{0, \dots, N-1\}, \qquad \begin{cases} p_{n+1}^k = (1-p_1)\sum_{l=0}^k p_n^l p_n^{k-l} + p_1 \sum_{l=0}^k q_n^l p_n^{k-l} \\ q_{n+1}^k = (1-p_2) p_n^k + p_2 \sum_{l=0}^k q_n^l q_n^{k-l} \end{cases} \tag{4.8}$$

A partir du système (4.7), il est également possible de calculer récursivement l'espérance $E_n[s]$ (resp. $E_n[m]$) du nombre de lettres m dans un texte généré aléatoirement par $L[s]$ (resp. $L[m]$) après n étapes de production. Il faut pour cela utiliser le corollaire A.1.2 de l'annexe A.1 et nous obtenons :

$$\forall n \in \{0, \dots, N-1\}, \qquad \begin{cases} E_{n+1}[s] = 2(1-p_1)E_n[s] + p_1\left(E_n[s] + E_n[m]\right) \\ E_{n+1}[m] = (1-p_2)E_n[s] + 2p_2 E_n[m] \end{cases} \tag{4.9}$$

Une équation similaire peut être obtenue pour la variance.

Le système d'équations (4.7) peut également être obtenu à partir du corollaire 4.2.6 car nous sommes en présence de processus de branchement multitype, voir Kang and de Reffye (2007) et Loi and Cournède (2008). En effet, le vecteur de fonctions génératrices $\mathbf{\Psi_n^L}$ de W_n^L associé à $V = \{s, m\}$ vérifie alors :

$$\mathbf{\Psi_{n+1}^L} = \mathbf{\Psi_1^L} \circ \mathbf{\Psi_n^L}$$

avec :

$$\mathbf{\Psi_n^L}(z_1, z_2) = \left(\sum_{w\in W} (P^n)_{s,w} {z_1}^{c(s,m)} {z_2}^{c(w,m)}, \sum_{w\in W} (P^n)_{m,w} {z_1}^{c(w,s)} {z_2}^{c(w,m)} \right).$$

$\mathbf{\Psi_1^L}$ peut directement être calculée à partir de L :

$$\mathbf{\Psi_1^L}(z_1, z_2) = \left((1-p_1)z_1^2 + p_1 z_1 z_2, (1-p_2)z_1 + p_2 z_2^2\right).$$

Le lien avec les fonctions génératrices $\Psi_n^{L[s]}$ et $\Psi_n^{L[m]}$ est le suivant :

$$\mathbf{\Psi_n^L}(1, z) = (\Psi_n^{L[s]}(z), \Psi_n^{L[m]}(z))$$

ce qui nous permet de retrouver le résultat précédent. De façon générale, les deux méthodes sont équivalentes lorsque l'on traitre de mots composés uniquement d'une seule lettre.

4.4.2 Exemple complexe

Soit $V = \{s, m, d\}$ un alphabet. Soit $L = \langle W^+, \pi \rangle$ un F0L-système stochastique dont les composantes de la matrice de transition π sont toutes égales à 0 à l'exception de :

$$\pi_{s,mssd} = p \qquad \pi_{s,\epsilon} = 1 - p \qquad \pi_{m,m} = 1 \qquad \pi_{d,d} = 1$$

avec $p \in]0, 1[$. Le noyau de transition correspondant est noté P. Nous souhaitons calculer la distribution associée au nombre d'occurrences du mot dm dans un texte généré aléatoirement par le système composant $L[s]$ après N étapes de production. Afin de résoudre ce problème, nous voulons déterminer la fonction génératrice de $W_N^{L[s]}$ associée à dm :

$$\Psi_N^{[s]}(z) = \sum_{w \in W} (P^N)_{s,w} \;\; z^{c(w,dm)} = \sum_{k \in \mathbb{N}} \;\; \mathbb{P}\left(c(w_N^{L[s]}, dm) = k\right) \;\; z^k$$

avec $\left(w_n^{L[s]}\right)_{n \geq 0}$ la chaîne de Markov engendrée par $L[s]$. Trouvons maintenant une spécification itérative appropriée pour $\{W_n^{L[s]} \;\; | \;\; n \in \{0, \ldots, N\}\}$. Une première idée est d'utiliser la même approche que dans l'exemple précédent (on utilise d'abord le théorème de décomposition générale 4.3.2 et ensuite on applique le théorème d'indépendance d'évolution 4.3.3). On obtient alors :

$$\forall n \in \{0, \ldots, N-1\}, \qquad W_{n+1}^{L[s]} = \{(\epsilon, 1-p)\} + \{(\epsilon, p)\}.W_n^{L[m]}.W_n^{L[s]}.W_n^{L[s]}.W_n^{L[d]}.$$

Etant donné que $\pi_{m,m} = 1$ et $\pi_{d,d} = 1$, nous avons :

$$\forall n \in \mathbb{N}, \qquad W_n^{L[m]} = \{(m, 1)\} \qquad W_n^{L[d]} = \{(d, 1)\}. \tag{4.10}$$

Ainsi :

$$\forall n \in \{0, \ldots, N-1\}, \qquad W_{n+1}^{L[s]} = \{(\epsilon, 1-p)\} + \{(\epsilon, p)\}.\{(m, 1)\}.W_n^{L[s]}.W_n^{L[s]}.\{(d, 1)\}. \tag{4.11}$$

Cependant, cette dernière équation ne repose pas sur des constructions admissibles puisque la concaténation de $W_n^{L[s]}$ et de $W_n^{L[s]}$ est génératrice par rapport à dm (voir la définition 4.3.3). En effet, comme le mot $mssd$ peut être obtenu après une étape de production de $L[s]$, la concaténation $mssd.mssd = mssdmssd$ est alors un mot pondéré de $W_1^{L[s]}.W_1^{L[s]}$ avec un poids p^2 strictement positif. Or, nous avons $c(mssdmssd, dm) \neq c(mssd, dm) + c(mssd, dm)$. Par conséquent, il est nécessaire de trouver une autre décomposition afin de pouvoir utiliser les règles de transformation du théorème 4.3.1. Etant donné que le mot dm ne peut être créé que par la concaténation de deux lettres s (voir les règles de production du L-système), l'idée est d'écrire une décomposition impliquant $W_n^{L[ss]}$. Nous pouvons l'obtenir à partir de l'équation (4.11) et du théorème 4.3.3 :

$$\forall n \in \{0, \ldots, N-1\}, \qquad W_{n+1}^{L[s]} = \{(\epsilon, 1-p)\} + \{(\epsilon, p)\}.\{(m, 1)\}.W_n^{L[ss]}.\{(d, 1)\}. \tag{4.12}$$

Cependant, l'équation (4.12) n'est pas suffisante pour avoir une spécification. Il faut alors une équation récurrente pour $W_n^{L[ss]}$ mettant en évidence le mot dm dans la dé-

composition :

$$\begin{aligned} W_{n+1}^{L[ss]} =& W_{n+1}^{L[s]}.W_{n+1}^{L[s]} \\ =& \left[\{(\epsilon,1-p)\}+\{(\epsilon,p)\}.\{(m,1)\}.W_n^{L[ss]}.\{(d,1)\}\right].\left[\{(\epsilon,1-p)\}+\{(\epsilon,p)\}.\{(m,1)\}.W_n^{L[ss]}.\{(d,1)\}\right] \\ =& \{(\epsilon,(1-p)^2)\}+\{(m,2p(1-p))\}.W_n^{L[ss]}.\{(d,1)\}+\{(m,p^2)\}.W_n^{L[ss]}.\{(dm,1)\}.W_n^{L[ss]}.\{(d,1)\}. \end{aligned} \tag{4.13}$$

Les équations (4.12) et (4.13) forment bien cette fois une spécification itérative appropriée pour $\{W_n^{L[s]} \mid n \in \{0,\ldots,N\}\}$. En effet, il est facile de prouver par une récurrence immédiate sur n que tous les mots de $W_n^{L[ss]}$ commencent soit par s ou m et se finissent soit par s ou d. Ainsi, dans les équations (4.12) et (4.13), la concaténation de deux ensembles consécutifs de mots pondérés est toujours non génératrice par rapport au mot dm. Nous avons donc des constructions admissibles. En appliquant les règles de transformation, nous obtenons alors :

$$\forall n \in \{0,\ldots,N-1\}, \quad \begin{cases} \Psi_{n+1}^{L[s]}(z) = 1-p+p\Psi_n^{L[ss]}(z) \\ \Psi_{n+1}^{L[ss]}(z) = (1-p)^2+2p(1-p)\Psi_n^{L[ss]}(z)+p^2 z\left(\Psi_n^{L[ss]}(z)\right)^2 \end{cases} \tag{4.14}$$

avec $\Psi_0^{L[s]}(z)=1$ et $\Psi_0^{L[ss]}(z)=1$. Tout comme pour l'exemple de la section précédente, les coefficients de $\Psi_N^{L[s]}(z)$ peuvent facilement être extrait du système d'équations (4.14) en identifiant les coefficients des séries entières correspondantes. En faisant ainsi, ces coefficients peuvent être calculés récursivement. Il est également possible de calculer de façon récursive l'espérance et la variance du nombre de mots dm dans un texte généré aléatoirement par $L[s]$ après N étapes de production en utilisant le corollaire A.1.2 de l'annexe A.1.

Chapitre 5

Étude combinatoire des structures de plante

Dans le chapitre 3, nous avons présenté un modèle de développement stochastique (noté $\mathcal{S}$) pour la croissance des plantes. Ce modèle a fait l'objet d'une étude probabiliste : mise en évidence des différents processus stochastiques et leur modélisation, écriture des fonctions génératrices, étude de la finitude de la croissance. Dans ce chapitre, nous introduisons un cadre combinatoire pour ce modèle permettant d'approfondir l'étude probabiliste. La structure d'une plante est codée par un mot de Dyck. Le développement de la plante est représenté par un 0L-système stochastique (voir la section 5.1.3).

L'utilisation de grammaires formelles pour créer des structures de plante n'est pas une idée nouvelle (voir Smith (1984), Prusinkiewicz and Lindenmayer (1990), Françon (1990), Kurth and Sloboda (1997), Kang and de Reffye (2007) entre autres). Elles ont surtout été employées pour des besoins graphiques (Prusinkiewicz and Lindenmayer (1990), Kruszewski and Whitesides (1998) ...). Plus récemment, (Kang et al. (2007)) les ont utilisées pour fournir un cadre mathématique propice à l'étude de l'organogenèse stochastique du modèle de croissance GreenLab 2. Dans un tel contexte, la composition de la plante à un cycle donné est codée par un mot construit à partir d'un alphabet de symboles représentant les bourgeons et les phytomères de tout âge physiologique. L'introduction des fonctions génératrices associées au nombre de lettres permettent de calculer de façon récursive les moments à tout ordre de n'importe quel type d'organe dans la plante à un cycle donné. Dans ce chapitre, nous utilisons un codage par mot de Dyck pour représenter la structure d'une plante (voir Loi et al. (2009, 2010)). Ce codage présente l'avantage de conserver la topologie de la plante (contrairement au codage de Kang et al. (2007)). Il est alors possible d'étendre l'étude de la composition à celle de motifs (un enchaînement particulier d'organes, des organes placés à des positions particulières ...). Pour cela, nous utiliserons les résultats du chapitre précédent et en particulier la méthode symbolique établie pour des textes générés aléatoirement par des 0L-systèmes stochastiques.

Nous commençons le chapitre en introduisant le cadre combinatoire associé au modèle de développement stochastique $\mathcal{S}$, section 5.1. Ensuite, la section 5.2 traite de l'occurrence de motifs dans des structures de plante générées par $\mathcal{S}$ et montre comment

appliquer la méthode symbolique écrite dans le chapitre précédent dans un tel contexte. Le chapitre se termine par l'application des résultats obtenus à l'estimation des vecteurs de paramètres Θ_{org} et Θ_{dif} du modèle $\mathcal{S}$ à partir de données botaniques, voir section 5.3.

5.1 Cadre combinatoire

Cette section commence par quelques rappels de combinatoire. Nous montrons ensuite comment la structure d'une plante peut être codée par un mot de Dyck et comment le modèle de développement $\mathcal{S}$ peut être représenté par un 0L-système stochastique.

5.1.1 Quelques concepts de combinatoire

Nous rappelons quelques concepts de combinatoire indispensables pour la suite du chapitre (voir Riordan (2002) et Flajolet and Sedgewick (2009) pour plus de détails).

Définition 5.1.1 (Graphe) *Un graphe simple non-orienté G est un ensemble (S, A) avec :*

- *S l'ensemble des sommets de G ;*
- *$A \subset \mathcal{P}_2(S)$ un ensemble de paires d'éléments de V appelé ensemble des arêtes de G.*

$\mathcal{P}_2(S)$ désigne l'ensemble des parties de cardinalité 2 de S.

Pour un graphe simple non-orienté, on appelle chemin un ensemble consécutif d'arêtes reliant deux sommets entre eux. La longueur d'un chemin est le nombre d'arêtes le constituant.

Définition 5.1.2 (Arbre) *Un arbre est un graphe simple non-orienté $G = (S, A)$ qui est :*

- *acyclique : il n'existe aucun chemin commençant et terminant par le même sommet ;*
- *connexe : pour tout couple de sommets de S, il existe un chemin permettant de les relier.*

La figure 5.1 montre un exemple d'arbre. Concernant les arbres, les sommets S sont aussi appelés nœuds.

Définition 5.1.3 (Arbre enraciné) *Un arbre $G = (S, A)$ est dit enraciné s'il existe un nœud que l'on distingue spécifiquement des autres et que l'on appelle racine.*

Pour un arbre enraciné, on appelle profondeur d'un nœud la longueur du chemin reliant directement la racine au nœud en question. Lorsque l'arbre est enraciné, il devient un graphe orienté. Les arêtes de A deviennent des arcs : la paire de sommets définissant une arête devient ordonnée. Le sommet de départ de chaque arc est donné par le nœud le moins profond (et le sommet d'arrivée par le nœud le plus profond). Dans le cas d'un

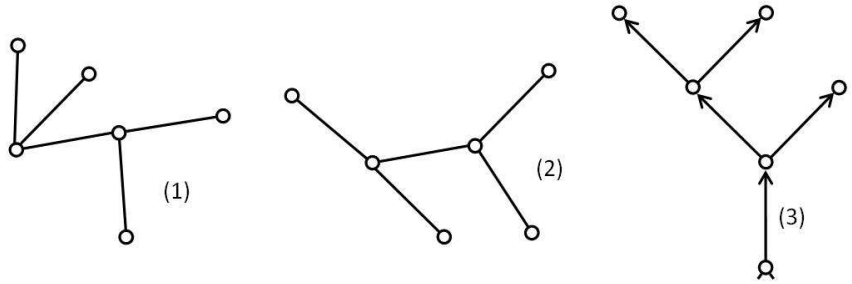

FIG. 5.1 – Exemples d'arbres. Si on les considère non orientés, les trois graphes représentent le même arbre. L'arbre (3) est un arbre enraciné. La racine est le nœud marqué par deux traits et sera toujours représentée par le nœud en bas de l'arbre. Les flèches indiquent le sens des arcs.

arbre, lorsque deux sommets sont reliés entre eux par un arc, le sommet de départ est appelé nœud parent du sommet d'arrivée (le sommet d'arrivé est appelé nœud fils du sommet de départ).

Définition 5.1.4 (Arbre planaire enraciné) *Un arbre enraciné est dit planaire si les sous-arbres provenant d'un même sommet sont ordonnés entre eux et représentés de gauche à droite.*

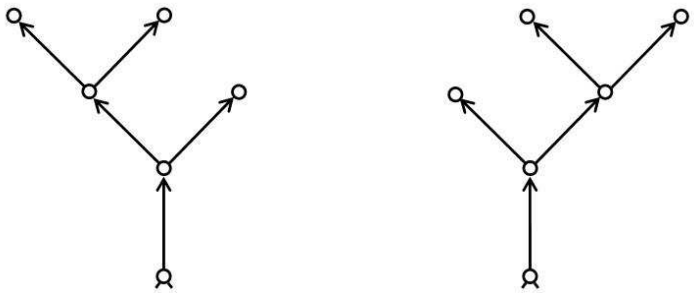

FIG. 5.2 – Exemples d'arbres planaires enracinés. Bien que ces deux graphes soient identiques en tant qu'arbres enracinés, ils deviennent deux objets distincts en tant qu'arbres planaires enracinés.

Il existe plusieurs façons de parcourir un arbre planaire enraciné. L'une des plus connues est le parcours en profondeur.

Définition 5.1.5 (Parcours en profondeur d'un arbre) *Un arbre planaire enraciné est parcouru en profondeur si :*

- *le premier nœud visité est la racine ;*
- *pour un nœud donné, l'ordre de visite des sous-arbres correspond à leur position de gauche à droite (la racine est placée en bas de l'arbre) ;*

- *une fois que tous les sous-arbres d'un nœud ont été visités, la visite se poursuit par le nœud parent ;*
- *le parcours de l'arbre est terminé lorsque tous les sous-arbres de la racine ont été visités.*

La figure 5.3 montre un exemple de parcours en profondeur. Le parcours en profondeur permet de coder les arbres planaires enracinés par un mot de Dyck (voir Knuth (1997 (3rd edition)).

Définition 5.1.6 (Mot de dyck) *Un mot w construit à partir de l'alphabet $V = \{z, z'\}$ est un mot de Dyck si :*

- *w comporte autant de lettres z que de lettres z' ;*
- *tout préfixe de w comporte un nombre de lettres z supérieur ou égal au nombre de lettres z'.*

Par exemple, $w_1 = zzz'z'$ et $w_2 = zzz'zz'z'$ sont des mots de Dyck. Nous avons alors :

Définition 5.1.7 (Mot de dyck associé à un arbre planaire enraciné) *Un arbre planaire enraciné peut être codé bijectivement par un mot de Dyck sur l'alphabet $V = \{z, z'\}$ de la façon suivante :*

- *l'arbre est visité suivant le parcours en profondeur ;*
- *chaque arête parcourue du nœud parent vers le nœud fils est codée par la lettre z ;*
- *chaque arête parcourue du nœud fils vers le nœud parent est codée par la lettre z'.*

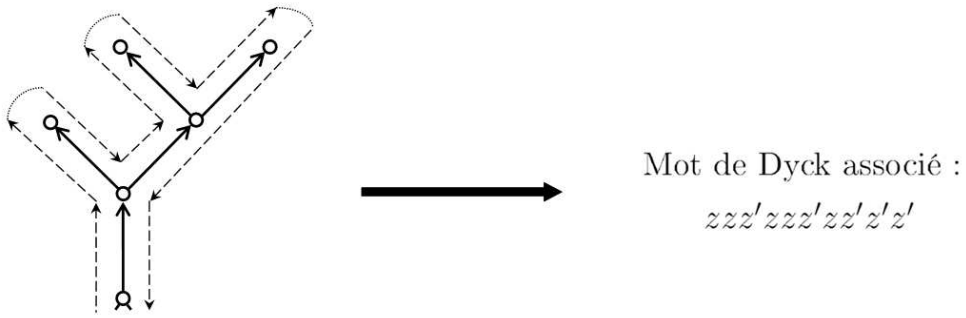

FIG. 5.3 – Mot de Dyck associé à un arbre planaire enraciné. Le parcours en profondeur est donné par les flèches en pointillé.

Il est possible d'étendre les notions précédentes lorsque des étiquettes peuvent être associées aux nœuds de l'arbre. Dans ce cas, l'arbre est dit étiqueté. Nous étendons alors la définition des mots de Dyck de la façon suivante :

Définition 5.1.8 (Mot de dyck généralisé) *Soit $\mathcal{L} = \{l_1, \ldots, l_m\}$ un ensemble de m étiquettes. Un mot w construit à partir de l'alphabet $V = \{z_{l_i}, z'_{l_i}\}_{i\in\{1,\ldots,m\}}$ est un mot de Dyck si :*

- *pour $i \in \{1, \ldots, m\}$, w comporte autant de lettres z_{l_i} que de lettres z'_{l_i} ;*
- *pour $i \in \{1, \ldots, m\}$, tout préfixe de w comporte un nombre de lettres z_{l_i} supérieur ou égal au nombre de lettres z'_{l_i}.*

Les mots $w_1 = z_{l_1} z_{l_1} z'_{l_1} z'_{l_1}$ et $w_1 = z_{l_1} z_{l_2} z_{l_1} z'_{l_2} z'_{l_1} z'_{l_1}$ sont des exemples de mots de Dyck sur l'alphabet $V = \{z_{l_1}, z'_{l_1}, z_{l_2}, z'_{l_2}\}$. Les arbres planaires enracinés et étiquetés peuvent également être codés par un mot de Dyck au sens de la définition 5.1.8 :

Définition 5.1.9 (Mot de Dyck associé à un arbre planaire enraciné étiqueté) *Soit $\mathcal{L} = \{l_1, \ldots, l_m\}$ un ensemble de m étiquettes. Un arbre planaire enraciné et étiqueté peut être codé bijectivement par un mot de Dyck sur l'alphabet $V = \{z_{l_i}, z'_{l_i}\}_{i\in\{1,\ldots,m\}}$ de la façon suivante :*

- *l'arbre est visité suivant le parcours en profondeur ;*
- *le mot de Dyck commence par la lettre z_l et se finit par la lettre z'_l avec $l \in \mathcal{L}$ l'étiquette de la racine.*
- *chaque arête parcourue du nœud parent vers le nœud fils est codée par la lettre z_{l_i} avec l_i l'étiquette du nœud fils ;*
- *chaque arête parcourue du nœud fils vers le nœud parent est codée par la lettre z'_{l_i} avec l_i l'étiquette du nœud fils ;*

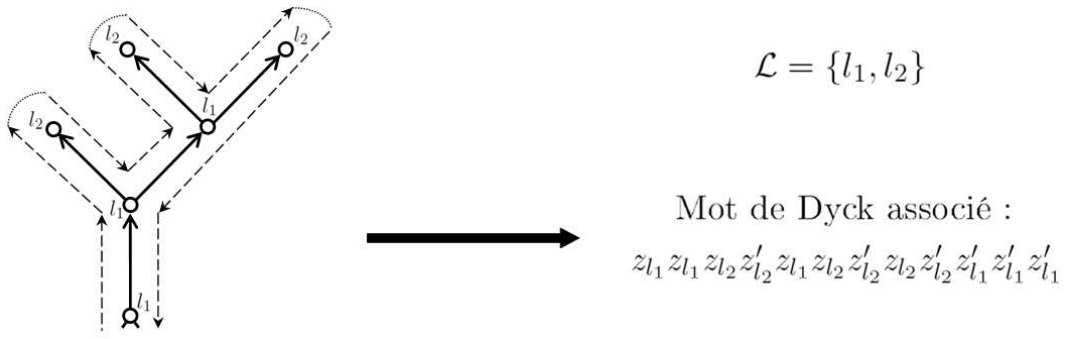

FIG. 5.4 – Mot de Dyck associé à un arbre planaire enraciné et étiqueté. L'ensemble des étiquettes est $\mathcal{L} = \{l_1, l_2\}$. A noter que le mot de Dyck associé à cette arbre est plus long que celui du cas non-étiqueté, cf figure 5.3. Ceci est dû à l'ajout de deux lettres permettant de coder l'étiquette de la racine et ainsi de conserver la bijectivité des représentations.

5.1.2 Codage de la structure d'une plante par un mot de Dyck

Dans cette section, nous montrons que toute structure de plante générée par le modèle de développement stochastique $\mathcal{S}$ peut être codée de façon bijective par un mot de Dyck (rappelons que, d'après la définition 2.1.5 du chapitre 2, une structure de classe physiologique $i \in \{1, \ldots, CP_m\}$ et d'âge $k \geq 0$ est la structure formée par l'ensemble des phytomères issus d'un bourgeon de classe physiologique i pendant k cycles de développement).

N.B. 5.1 Dans ce chapitre, nous faisons la distinction entre deux structures de plante dont l'une est l'image chirale de l'autre. Ainsi, les deux structures de la figure 5.5 sont considérées comme étant des objets distincts bien que, du point de vue 3D, elles soient identiques. Le passage de la 2D à la 3D se fait au niveau de la calibration des automates donnant les règles de production du développement (voir la section 3.1.3). En effet, supposons par exemple que les deux plantes 2D de la figure 5.5 aient respectivement une probabilité d'occurrence p_1 et p_2. Si on les considère comme structures 3D, alors leur probabilité d'occurrence sera $p_1 + p_2$.

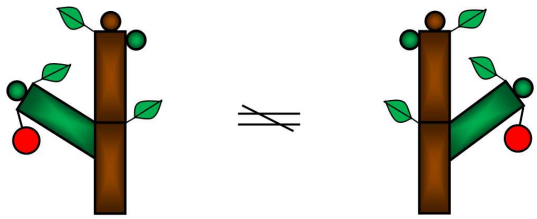

FIG. 5.5 – Deux configurations 2D possibles pour une même structure 3D. Dans ce chapitre, les deux plantes ci-dessus sont deux objets différents.

Afin de faire le lien entre structures de plante et mots de Dyck, nous montrons auparavant la proposition suivante :

Proposition 5.1.1 *Toute structure de plante générée par le modèle de développement stochastique $\mathcal{S}$ peut être représentée bijectivement par un arbre planaire enraciné et étiqueté.*

Preuve Soit $i \in \{1, \ldots, CP_m\}$ et $k \geq 0$. Pour prouver la proposition, nous allons construire l'arbre planaire enraciné et étiqueté associé à une structure de CP i et d'âge k. Soit $\mathcal{L} = \mathcal{B} \cup \mathcal{O}$ l'ensemble des étiquettes avec $\mathcal{B}$ et $\mathcal{O}$ les ensembles de symboles décrivant respectivement les bourgeons caractérisés par leur classe physiologique et les autres organes composant la structure de la plante (voir la section 2.1.6). Si $k = 0$, alors la structure se réduit à un bourgeon de CP i. Celle-ci peut donc être représentée par un arbre planaire enraciné et étiqueté composé de deux nœuds reliés entre eux par un arc : le nœud fils représente le bourgeon et porte l'étiquette b_i et le nœud parent est la racine à laquelle on attribue l'étiquette r. Si $k > 0$, cette structure peut alors être représentée par l'arbre planaire enraciné et étiqueté défini de la façon suivante :

- tout organe de type $l \in \mathcal{L}$ est représenté par un nœud dont l'étiquette est l ;
- la racine est le nœud parent du nœud représentant l'entrenœud de CP i situé à la base de la structure et porte l'étiquette r ;
- deux organes connectés sont représentés par deux nœuds reliés entre eux par un arc. L'organe porteur est représenté par le nœud parent et l'organe porté par le nœud fils ;
- la position des nœuds et des arcs de l'arbre respecte l'agencement des organes donné par la structure.

L'arbre planaire enraciné et étiqueté ainsi défini représente bien de façon bijective la structure de plante.

□

Une plante d'âge k est une structure de CP 1 et d'âge k. Elle peut donc être représentée bijectivement par un arbre planaire enraciné et étiqueté. Dans ce cas, le nœud racine de l'arbre planaire peut également représenter l'organe racine de la plante.

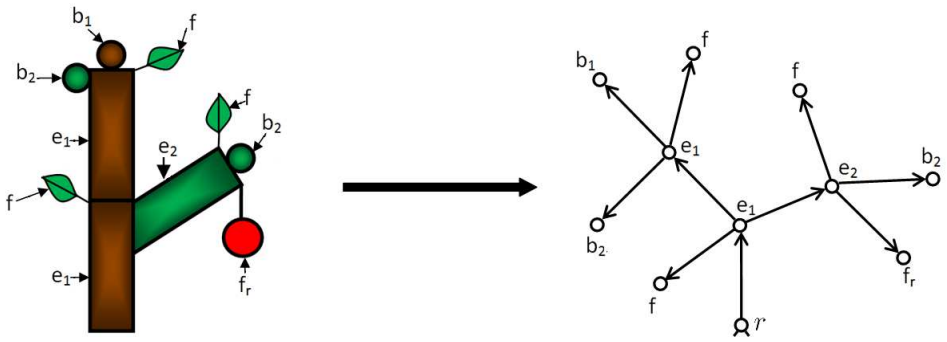

FIG. 5.6 – Arbre planaire enraciné et étiqueté associé à une plante. Les pétioles p et les limbes l sont représentés par un unique symbole f (pour feuille).

N.B. 5.2 Godin and Caraglio (1998) avaient déjà fait le lien entre structures de plante et arbres planaires. Leur représentation est différente de celle présentée dans la thèse et est utilisée entre autre pour des besoins de simulation.

Un corollaire immédiat est le suivant :

Corollaire 5.1.2 *Toute structure de plante générée par le modèle de développement stochastique $\mathcal{S}$ peut être codée bijectivement par un mot de Dyck sur l'alphabet $V = \{z_l, z'_l\}_{l \in \mathcal{B} \cup \mathcal{O}}$.*

Preuve Le résultat découle du fait que tout arbre planaire enraciné et étiqueté peut être codé bijectivement par un mot de Dyck, cf définition 5.1.9.

□

Afin de simplifier l'écriture des mots de Dyck associés aux structures de plante, nous adoptons les conventions de notation suivantes :

- les lettres z_r et z'_r sont retirées du mot de Dyck : elles se situent toujours respectivement en première et dernière position et n'apportent aucune information sur la structure de la plante ;
- chaque lettre $z_l \in V$ (resp. $z'_l \in V$) est remplacée par l'étiquette l (resp. l') qui lui est associée ;
- pour les organes de type $l \in \mathcal{B} \cup \{l, p, f_r, f_l\}$ (resp. les bourgeons, limbes, pétioles, fruits et fleurs), la lettre z_l est toujours suivie de la lettre z'_l. Dans ce cas, la séquence $z_l z'_l$ sera remplacée par z_l. Les seuls organes qui ne sont pas concernés par cette convention sont les entrenœuds de CP i noté e_i, $i \in \{1, \ldots, CP_m\}$.

La figure 5.7 donne un exemple de codage avec un mot de Dyck et le mot de Dyck simplifié qui en résulte suite aux conventions de notation.

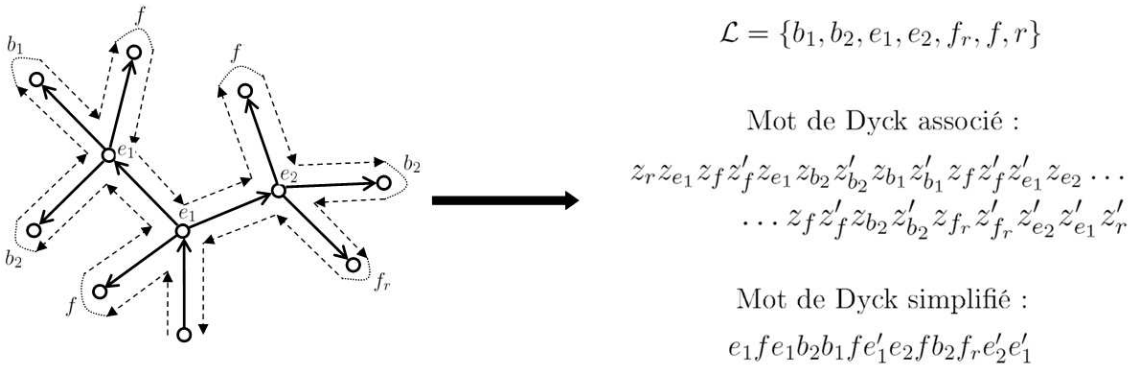

FIG. 5.7 – Mots de Dyck associés à une plante. Le parcours en profondeur est donné par les traits en pointillé.

Pour la suite du chapitre, V désigne l'alphabet des symboles d'organe permettant de coder par un mot de Dyck la structure d'une plante générée par $\mathcal{S}$:

$$V = \mathcal{B} \cup \{e_i, e'_i\}_{i \in \{1,\ldots,CP_m\}} \cup \{l, p, f_r, f_l\}.$$

L'ensemble des mots finis construits à partir de V est noté W et l'ensemble des mots non vides W^+. Le mot vide sera désigné par ϵ.

5.1.3 Développement et L-systèmes stochastiques

Dans la section précédente, nous avons vu que toute structure de plante à un cycle de développement $n \in \mathbb{N}$ donné peut être codée bijectivement par un mot de Dyck w_n. Dans cette section, nous nous intéressons à la dynamique d'évolution $(w_n)_{n \geq 0}$ pour une structure générée suivant le modèle $\mathcal{S}$ cycle de développement après cycle de développement. Autrement dit, nous cherchons à intégrer l'organogenèse stochastique et la différenciation dans le cadre combinatoire initié dans la section précédente. L'outil mathématique approprié est alors le 0L-système stochastique (voir le chapitre 4). En effet,

les règles de production associées à $\mathcal{S}$ peuvent être représentées de façon synthétique par une matrice de transition π de V dans W de la façon suivante :

- pour $b \in \mathcal{B}$ et $w \in W$, $\pi_{b,w}$ donne la probabilité que le bourgeon de type b évolue vers la structure de plante w (probabilité donnée par les règles de production). On peut noter que, si w n'est pas un mot de Dyck (en tenant compte des conventions de notation détaillées dans la section précédente), alors $\pi_{b,w} = 0$ car toute structure de plante peut être codée bijectivement par un mot de Dyck, cf corollaire 5.1.2.
- pour $o \in V \setminus \mathcal{B}$ et $w \in W$, $\pi_{o,w} = \delta_o(w)$ avec δ_o le symbole de Kronecker centré en o. En effet, tout organe $o \in V \setminus \mathcal{B}$ reste inchangé par les règles de production de $\mathcal{S}$.

N.B. 5.3 Notons que la matrice de transition π ainsi définie ne prend pas en compte la sénescence des feuilles. Autrement dit, à partir du moment où une feuille est rajoutée à la structure de la plante, celle-ci reste en permanence. Pour corriger ce problème, il faudrait ajouter un indice au symbole f (resp. l et p si la distinction est faite entre limbe et pétiole) traduisant l'âge de l'organe. Cette représentation complique alors l'étude de la structure de la plante et ne sera pas retenue pour cette thèse.

Prenons par exemple les règles de production données par la figure 5.8. Dans ce cas, l'alphabet est $V = \{b_1, b_2, e_1, e_2, f_r, r\}$. Toutes les composantes de la matrice de transition π associée au modèle $\mathcal{S}$ sont nulles à l'exception de :

$$\pi_{b_1,b_1} = 1 - p_a(1) \qquad \pi_{b_1,e_1 b_1 b_2 e_1'} = p_a(1) \qquad \pi_{b_2,e_2 b_2 f_r e_2'} = 1$$

pour les bourgeons et de :

$$\pi_{e_1,e_1} = 1 \qquad \pi_{e_1',e_1'} = 1 \qquad \pi_{e_2,e_2} = 1 \qquad \pi_{e_2',e_2'} = 1 \qquad \pi_{f_r,f_r} = 1$$

pour les autres organes.

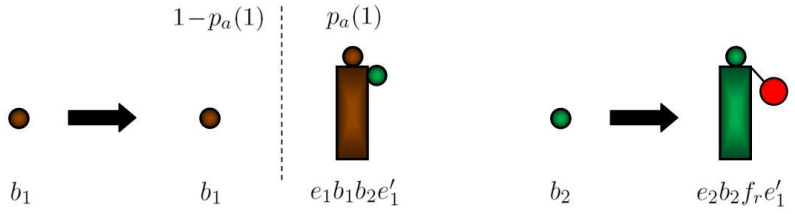

FIG. 5.8 – Règles de production et 0L-système stochastique.

On a alors la proposition suivante :

Proposition 5.1.3 *Soit π la matrice de transition de V dans W donnant les règles de production de $\mathcal{S}$. Pour tout $i \in \{1, \ldots, CP_m\}$, la suite de mots de Dyck associée à l'évolution d'une structure de classe physiologique i générée par le modèle S coïncide avec celle générée par le 0L-système stochastique $L = \langle b_i, \pi \rangle$. Dans ce cas, on dit que L représente le développement de la structure de classe physiologique i générée par $\mathcal{S}$.*

Preuve Soit $i \in 1, \ldots, CP_m$. Notons w_n le mot de Dyck représentant une structure de CP i et d'âge n. Le mot w_0 correspond à la structure d'âge 0. Nous avons donc $w_0 = b_i$. La structure au cycle $n+1$ est obtenue en remplaçant chaque bourgeon de la structure au cycle n par l'une de ses évolutions selon les règles de production de $\mathcal{S}$. Autrement dit, w_{n+1} est obtenu en remplaçant chaque lettre b de w_n par un mot de Dyck w avec une probabilité $\pi_{b,w}$ (les lettres $o \in V \setminus \mathcal{B}$ de w_n associées aux organes qui ne sont pas des bourgeons restent inchangées dans w_{n+1}). Le processus à temps discret $(w_n)_{n\geq 0}$ est donc une chaîne de Markov. En suivant les règles de production de $\mathcal{S}$, les probabilités de transition sont données par :

$$\forall (x,y) \in W \times W, \qquad \mathbb{P}(w_{n+1}^L = y | w_n^L = x) = \sum_{\substack{(y_1,y_2,\ldots,y_n)\in W^n,\\ y_1.y_2\ldots y_n = y}} \prod_{i=1}^{n} \pi_{x_i,y_i}$$

avec $x = x_1 x_2 \ldots x_n$ et $x_i \in V$ pour tout $i \in \{1, \ldots, n\}$. Ainsi, le processus à temps discret $(w_n)_{n\geq 0}$ possède le même espace d'états W, le même noyau de transition et la même distribution initiale que la chaîne de Markov $\left(w_n^L\right)_{n\geq 0}$ générée par le 0L-système stochastique $L = \langle b_i, \pi \rangle$ ce qui prouve la proposition.

□

A partir de la proposition précédente, il est naturel de définir le concept de F0L-système stochastique associé au modèle $\mathcal{S}$:

Définition 5.1.10 (F0L-système associé au modèle $\mathcal{S}$) *Soit π la matrice de transition de V dans W donnant les règles de production de $\mathcal{S}$. La grammaire formelle $L = \langle W^+, \pi \rangle$ est appelée F0L-système stochastique associé au modèle de développement $\mathcal{S}$.*

Ainsi, pour $b \in \mathcal{B}$, le système composant $L[b]$ caractérise l'organogenèse d'une structure de plante initiée par un bourgeon de type b. Par la suite, nous utiliserons les mêmes notations que dans le chapitre 4. Ainsi, pour tout 0L-système stochastique $L = \langle b, \pi \rangle$ de noyau de transition P et $b \in \mathcal{B}$, W_n^L désignera l'ensemble des mots pondérés possibles générés par L après n étapes de production :

$$W_n^L = \{(w, (P^n)_{a,w}) \ \mid \ w \in W\}.$$

Du point de vue botanique, W_n^L peut être vue comme l'ensemble des structures de plante initiées par un bourgeon de type b et générées par le modèle de développement stochastique $\mathcal{S}$ après n cycles de développement. Pour $w \in W$, $(P^n)_{a,w}$ est la probabilité de réalisation d'une telle structure de plante.

5.2 Occurrence de motifs dans la structure d'une plante

5.2.1 Principe

Dans le chapitre 3, nous avons effectué une étude probabiliste du modèle de développement stochastique $\mathcal{S}$. Nous avons entre autre mis en évidence que le processus

$(N_n)_{n\in\{0,\dots,T\}}$, pour $T \in \mathbb{N}$, est un processus de branchement de Galton-Watson multi-type, voir le théorème 3.3.3 du chapitre 3. Nous rappelons que N_n est le vecteur donnant la composition de la plante au début du cycle de développement n (avant l'organogenèse) :

$$N_n = (N_n^a)_{a\in\mathcal{B}\cup\mathcal{O}}, \qquad n \geq 0,$$

avec N_n^a le nombre total d'organes de type a dans la plante au début du cycle de développement n. Nous avons également écrit une relation de récurrence pour la fonction génératrice $\Psi_n^{dev}[b_1]$ associée au vecteur aléatoire N_n, cf corollaire 3.4.4. Grâce à cette relation, il est possible d'obtenir de façon récursive la distribution de probabilité et les moments à tout ordre du nombre d'organes de tout type dans une plante à un cycle de développement n donné.

Cette approche trouve cependant ses limites quand on s'intéresse à des répétitions de motifs dans la topologie de la plante (*i.e.* la façon dont les organes sont reliés les uns aux autres). Ici, le terme motif peut désigner un agencement particulier d'organes (par exemple, les structures en Y de la section 5.2.3), une sous-structure précise de la plante ou encore des organes occupant une position particulière (cf section 5.2.4 avec les apex). Dans l'étude probabiliste du chapitre 3, la topologie de la plante n'était pas prise en compte. Ce qui nous intéressait alors était la composition de la plante à un cycle donné (c'est-à-dire le nombre d'organes de chaque type). Cependant, l'étude de motifs nécessite de conserver l'aspect structurel de la plante. Le cadre combinatoire détaillé dans la section 5.1 est plus adapté au problème grâce à l'utilisation des mots de Dyck qui servent à coder bijectivement la structure de la plante.

Soit π la matrice de transition donnant les règles de production de $\mathcal{S}$. Soit w_N le mot de Dyck codant la structure d'une plante générée par $\mathcal{S}$ après N cycles de développement. Dans le cadre combinatoire détaillé dans ce chapitre, un motif est en fait une chaîne de caractères u pris dans l'alphabet V. Déterminer la distribution (ou les moments) associée au nombre d'occurrences du motif en question dans une plante générée par $\mathcal{S}$ après N cycles de développement est en fait équivalent à déterminer la distribution (ou les moments) associée au nombre d'occurrences du mot u dans le mot aléatoire w_N. Le problème se résume donc à l'étude d'occurrences du mot u dans un texte généré aléatoirement par le 0L-système stochastique $L = \langle b_1, \pi\rangle$ après N étapes de production. La méthode symbolique développée dans le chapitre 4 précédent permet de résoudre ce problème dans certains cas. En effet, pour N donné, la distribution d'intérêt est obtenue à partir des coefficients de la fonction génératrice Ψ_N^L de L associée à $\{u\}$ après N étapes de production et mise sous la forme d'une série entière :

$$\Psi_N^L(z) = \sum_{w\in W} (P^N)_{b_1,w}\, z^{c(w,u)} = \sum_{k\in\mathbb{N}} \mathbb{P}\left(c(W_N^L, u) = k\right) z^k$$

avec c la fonction de comptage donnée par la définition 4.1.6 du chapitre 4, P le noyau de Markov associé à L et $\left(W_n^L\right)_{n\geq 0}$ la chaîne de Markov engendrée par L. Grâce à la méthode symbolique, nous obtenons un ensemble de relations de récurrence faisant intervenir les fonctions génératrices Ψ_k^L avec $k \in \{0,\dots,N\}$. Il est alors possible d'extraire la distribution du nombre de motifs ou encore les moments associés.

Le reste de la section 5.2 est dédié à la mise en œuvre de la méthode symbolique au travers de trois exemples. Les deux premiers sont très fortement liés aux exemples d'application de la section 4.4. A chaque fois, l'objectif est d'obtenir les relations de récurrence vérifiées par les fonctions génératrices d'intérêt.

5.2.2 Distribution du nombre de fruits

Considérons le modèle de développement $\mathcal{S}$ suivant :
- La classe physiologique maximale est $CP_m = 2$;
- Un entrenœud de CP 1 porte un bourgeon apical de CP 1, un bourgeon axillaire de CP 2 et une feuille ;
- Un entrenœud de CP 2 porte un bourgeon apical de CP 2, une feuille et un fruit ;
- Un bourgeon de CP 1 est actif et produit un nouveau phytomère avec une probabilité $p_a(1)$ ou bien il est dormant avec une probabilité $1-p_a(1)$;
- Un bourgeon de CP 2 produit toujours un nouveau phytomère à chaque cycle de développement.

Les règles de production de $\mathcal{S}$ sont illustrées par la figure 5.8. Nous souhaitons déterminer la distribution associée au nombre de fruits dans une plante générée aléatoirement par $\mathcal{S}$ après N cycles de développement. Etant donné que nous ne nous intéressons pas aux feuilles et que celles-ci n'interviennent pas dans la mise en place de l'architecture de la plante, nous les omettons pour la suite de l'étude. Les ensembles de symboles associés aux bourgeons et aux autres organes de la plante sont alors donnés respectivement par $\mathcal{B} = \{b_1, b_2\}$ et $\mathcal{O} = \{e_1, e_2, f_r\}$. En tenant compte des conventions de notations pour les mots de Dyck associés aux structures de plante (voir section 5.1.2), l'alphabet est donc donné par :

$$V = \{b_1, b_2, e_1, e'_1, e_2, e'_2, f_r\}.$$

La matrice de transition π associée aux règles de production de $\mathcal{S}$ est la matrice de V dans W dont toutes les composantes sont nulles à l'exception de :

$$\pi_{b_1,b_1} = 1-p_a(1) \qquad \pi_{b_1,e_1b_1b_2e'_1} = p_a(1) \qquad \pi_{b_2,e_2b_2f_re'_2} = 1$$

pour les bourgeons et de :

$$\pi_{e_1,e_1} = 1 \qquad \pi_{e'_1,e'_1} = 1 \qquad \pi_{e_2,e_2} = 1 \qquad \pi_{e'_2,e'_2} = 1 \qquad \pi_{f_r,f_r} = 1$$

pour les autres organes. Soit $L = \langle W^+, \pi \rangle$ le F0L-système stochastique associé au modèle de développement $\mathcal{S}$. Déterminer la distribution associée au nombre de fruits dans une plante générée aléatoirement par $\mathcal{S}$ après N cycles de développement est équivalent à déterminer la distribution associée au nombre d'occurrences du mot f_r dans un texte généré aléatoirement par le système composant $L[b_1]$ après N étapes de production. Pour résoudre le problème, nous souhaitons calculer $\Psi_N^{L[b_1]}$, la fonction génératrice de $L[b_1]$ associée à $\{f_r\}$ après N étapes de production :

$$\Psi_N^{L[b_1]}(z) = \sum_{w \in W} (P^N)_{b_1,w}\, z^{c(w,f_r)} = \sum_{k \in \mathbb{N}} \mathbb{P}\left(c(W_N^{L[b_1]}, f_r) = k\right) z^k$$

avec P le noyau de Markov associé à L et $\left(W_n^{L[b_1]}\right)_{n\geq 0}$ la chaîne de Markov engendrée par $L[b_1]$. $\Psi_N^L(z)$ peut être déterminée récursivement en utilisant la méthode symbolique du chapitre 4. La première étape consiste à trouver une spécification itérative appropriée pour $\{W_n^{L[b_1]} \mid n \in \{0,\ldots,N\}\}$. Le problème rencontré ici est très proche de celui traité dans la section 4.4.1 à propos de l'occurrence de mots composés d'une seule lettre. Les mêmes techniques peuvent donc être employées. En utilisant d'abord le théorème de décomposition générale 4.3.2 et ensuite le théorème d'indépendance d'évolution 4.3.3, il vient :

$$\forall n \in \{0,\ldots,N-1\}, \qquad W_{n+1}^{L[b_1]} = \{(\epsilon, 1-p_a(1))\}.W_n^{L[b_1]} + \{(\epsilon, p_a(1))\}.W_n^{L[e_1]}.W_n^{L[b_1]}.W_n^{L[b_2]}.W_n^{L[e_1']} \tag{5.1}$$

Il nous faut également une équation combinatoire pour $W_n^{L[b_2]}$ avec $n \in \{0,\ldots,N\}$. Nous procédons de la même façon :

$$\forall n \in \{0,\ldots,N-1\}, \qquad W_{n+1}^{L[b_2]} = W_n^{L[e_2]}.W_n^{L[b_2]}.W_n^{L[f_r]}.W_n^{L[e_2']}. \tag{5.2}$$

D'après les règles de production données par π, il est immédiat que :

$$n \in \{0,\ldots,N\}, \quad a \in \{e_1, e_1', e_2, e_2', f_r\}, \qquad W_n^{L[a]} = \{(a,1)\}.$$

En remplaçant dans les équations (5.1) et (5.2), nous obtenons :

$$\forall n \in \{0,\ldots,N-1\}, \qquad \begin{cases} W_{n+1}^{L[b_1]} = \{(\epsilon, 1-p_a(1))\}.W_n^{L[b_1]} + \{(e_1, p_a(1))\}.W_n^{L[b_1]}.W_n^{L[b_2]}.\{(e_1',1)\} \\ W_{n+1}^{L[b_2]} = \{(e_2,1)\}.W_n^{L[b_2]}.\{(f_r e_2', 1)\} \end{cases}$$

Le système d'équations combinatoires précédent est bien une spécification itérative adéquate pour $\{W_n^{L[b_1]} \mid n \in \{0,\ldots,N\}\}$ puisque toutes les concaténations sont non-génératrices par rapport à f_r (voir la définition 4.3.3). Il est donc possible d'utiliser les règles de transformation du théorème 4.3.1 et on obtient le système d'équations récurrentes suivant :

$$\forall n \in \{0,\ldots,N-1\}, \qquad \begin{cases} \Psi_{n+1}^{L[b_1]}(z) = (1-p_a(1))\Psi_n^{L[b_1]}(z) + p_a(1)\Psi_n^{L[b_1]}(z)\Psi_n^{L[b_2]}(z) \\ \Psi_{n+1}^{L[b_2]}(z) = z\Psi_n^{L[b_2]}(z) \end{cases} \tag{5.3}$$

avec $\Psi_0^{L[b_1]}(z) = 1$ et $\Psi_0^{L[b_2]}(z) = 1$. Il s'en suit immédiatement :

$$\Psi_n^{L[b_2]}(z) = z^n, \qquad n \geq 0$$

et donc :

$$\Psi_{n+1}^{L[b_1]}(z) = \Psi_n^{L[b_1]}(z)\left(1 - p_a(1) + p_a(1)z^n\right) = \prod_{k=1}^{n}\left(1 - p_a(1) + p_a(1)z^k\right) \tag{5.4}$$

Soit $p_n^k = \mathbb{P}\left(c(W_n^{L[b_1]}, f_r) = k\right)$ la probabilité d'avoir k fruits dans une plante d'âge n. p_n^k représente le coefficient devant z^k dans le développement en série entière de $\Psi_n^{L[b_1]}$. Grâce

à l'équation (5.4), il est possible de déterminer récursivement la distribution $\left\{p_N^k\right\}_{k\in\mathbb{N}}$ en identifiant les coefficients des fonctions génératrices vues comme séries entières :

$$\forall n \in \{0,\dots,N-1\},\quad k\geq 0, \qquad p_{n+1}^k = (1-p_a(1))p_n^k + \mathbb{1}_{\mathbb{N}}(k-n)p_a(1)p_k^{n-k}.$$

avec $\mathbb{1}_{\mathbb{N}}$ l'indicatrice des entiers naturels. En dérivant l'équation (5.4) et en prenant $z=1$ (voir le corollaire A.1.2 de l'annexe A.1), nous obtenons l'espérance du nombre de fruits dans une plante d'âge N :

$$\mathbb{E}\left[c(W_N^{L[b_1]}, f_r)\right] = \frac{N(N-1)}{2}p_a(1), \qquad N\geq 1.$$

Notons que si $\hat{E}_N$ désigne la distribution empirique (obtenue à partir d'un ensemble de vraies plantes) du nombre de fruits dans une plante d'âge N, nous pouvons fournir une estimation de la probabilité $p_a(1)$ que nous notons $\hat{p}_a(1)$ par :

$$\hat{p}_a(1) = \hat{E}_N \frac{2}{N(N-1)}.$$

La section 5.3 consacrée à l'estimation des paramètres de $\mathcal{S}$ utilise en partie cette technique.

N.B. 5.4 Le système d'équations (5.3) aurait pu être obtenu directement en appliquant le théorème de composition des fonctions génératrices associées à N_n (voir corollaire 3.4.4 du chapitre 3) et en affectant la valeur 1 pour toutes les variables à l'exception de e_2. En effet, dans cet exemple, le nombre de fruits correspond au nombre d'entre-nœuds de CP 2.

5.2.3 Distribution du nombre de structures en Y

Considérons le modèle $\mathcal{S}$ défini de la façon suivante :
- La classe physiologique maximale est $CP_m = 1$;
- Un entrenœud de CP 1 porte deux bourgeons axillaires de CP 1, aucun bourgeon apical et une feuille ;
- Un bourgeon de CP 1 est actif et produit un nouveau phytomère avec une probabilité $p_s(1)$ ou bien il meurt avec une probabilité $1-p_s(1)$.

Les règles de production de $\mathcal{S}$ sont illustrées par la figure 5.9.

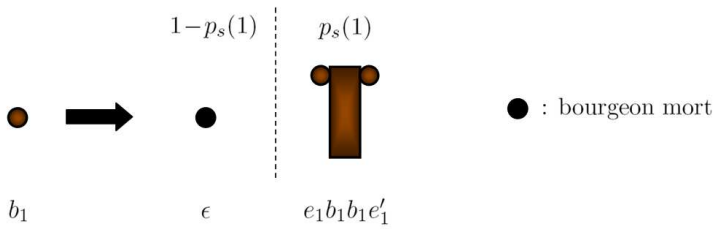

FIG. 5.9 – Règles de production avec probabilité de survie.

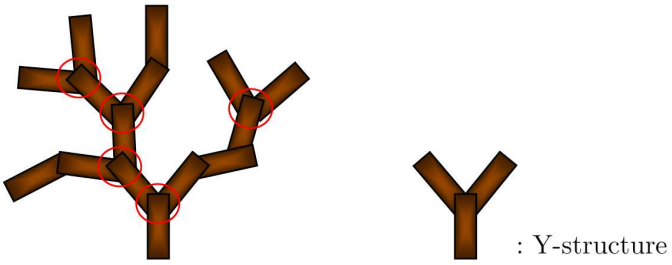

FIG. 5.10 – Exemples de structures en Y.

Nous souhaitons déterminer la distribution associée au nombre de structures en Y dans une plante générée aléatoirement par $\mathcal{S}$ après N cycles de développement. On appelle structure en Y la structure formée par un phytomère et les deux phytomères latéraux qu'il porte, cf figure 5.10.

Etant donné que nous ne nous intéressons pas aux feuilles et que celles-ci n'interviennent pas dans la mise en place de l'architecture de la plante, nous les omettons une fois de plus pour la suite de l'étude. Les ensembles de symboles associés aux bourgeons et aux autres organes de la plante sont alors donnés respectivement par $\mathcal{B} = \{b_1\}$ et $\mathcal{O} = \{e_1\}$. En tenant compte des conventions de notations pour les mots de Dyck associés aux structures de plante (voir section 5.1.2), l'alphabet est donc donné par :

$$V = \{b_1, e_1, e_1'\}.$$

La matrice de transition π associée aux règles de production de S est la matrice de V dans W dont toutes les composantes sont nulles à l'exception de :

$$\pi_{b_1,\epsilon} = 1 - p_s(1) \qquad \pi_{b_1,e_1b_1b_1e_1'} = p_s(1) \qquad \pi_{e_1,e_1} = 1 \qquad \pi_{e_1',e_1'} = 1$$

Soit $L = \langle W^+, \pi \rangle$ le F0L-système stochastique associé au modèle de développement $\mathcal{S}$. Déterminer la distribution associée au nombre de structures en Y dans une plante générée aléatoirement par $\mathcal{S}$ après N cycles de développement est équivalent à déterminer la distribution associée au nombre d'occurrences du mot $e_1'e_1$ dans un texte généré aléatoirement par le système composant $L[b_1]$ après N étapes de production. Pour résoudre le problème, nous souhaitons calculer $\Psi_N^{L[b_1]}$, la fonction génératrice de $L[b_1]$ associée à $\{e_1'e_1\}$ après N étapes de production :

$$\Psi_N^{L[b_1]}(z) = \sum_{w \in W} (P^N)_{b_1,w}\, z^{c(w,e_1'e_1)} = \sum_{k \in \mathbb{N}} \mathbb{P}\left(c(W_N^{L[b_1]}, e_1'e_1) = k\right) z^k$$

avec P le noyau de Markov associé à L et $\left(W_n^{L[b_1]}\right)_{n \geq 0}$ la chaîne de Markov engendrée par $L[b_1]$. Tout comme dans l'exemple précédent, $\Psi_N^{L[b_1]}(z)$ peut être déterminée récursivement en utilisant la méthode symbolique du chapitre 4. Le problème a déjà été traité dans la section 4.4.2. Il suffit de remplacer les lettres s, m et d respectivement par b_1, e_1 et e_1'. Une spécification itérative adéquate pour $\{W_n^{L[b_1]} \mid n \in \{0, \ldots, N\}\}$ est alors

donnée par :

$$\forall n \in \{0,\dots,N-1\},$$
$$\begin{cases} W_{n+1}^{L[b_1]} = \{(\epsilon, 1-p_s(1))\} + \{(e_1, p_s(1))\}.W_n^{L[b_1b_1]}.\{(e_1',1)\} \\ W_{n+1}^{L[b_1b_1]} = \{(\epsilon,(1-p_s(1))^2)\} + \{(e_1, 2p_s(1)(1-p_s(1)))\}.W_n^{L[b_1b_1]}.\{(e_1',1)\} \\ \qquad\qquad +\{(e_1,p_s(1)^2)\}.W_n^{L[b_1b_1]}.\{(e_1'e_1,1)\}.W_n^{L[b_1b_1]}.\{(e_1',1)\} \end{cases} \tag{5.5}$$

En utilisant les règles de transformation du théorème 4.3.1, nous obtenons le système d'équations récurrentes pour $\Psi_n^{L[b_1]}$ pour $n \in \{0,\dots,N\}$:

$$\forall n \in \{0,\dots,N-1\}, \quad \begin{cases} \Psi_{n+1}^{L[b_1]}(z) = 1 - p_s(1) + p_s(1)\Psi_n^{L[b_1b_1]}(z) \\ \Psi_{n+1}^{L[b_1b_1]}(z) = (1-p_s(1))^2 + 2p_s(1)(1-p_s(1))\Psi_n^{L[b_1b_1]}(z) + p_s(1)^2 z\left(\Psi_n^{L[b_1b_1]}(z)\right)^2 \end{cases} \tag{5.6}$$

avec $\Psi_0^{L[b_1]}(z) = 1$ et $\Psi_0^{L[b_1b_1]}(z) = 1$. On en déduit :

$$\Psi_n^{L[b_1b_1]}(z) = \frac{p_s(1)-1}{p_s(1)} + \frac{1}{p_s(1)}\Psi_{n+1}^{L[b_1]}(z)$$

et donc en réinjectant dans la deuxième équation du système (5.6) :

$$\forall n\{0,\dots,N-1\},$$
$$\Psi_{n+1}^{L[b_1]}(z) = 1 - p_s(1) + p_s(1)(1-p_s(1)) + 2p_s(1)\,(1-p_s(1))\,\Psi_n^{L[b_1]}(z) + p_s(1)z\left(p_s(1) - 1 + \Psi_n^{L[b_1]}(z)\right)^2 \tag{5.7}$$

avec $\Psi_1^{L[b_1]}(z) = 1$. Soit $p_n^k = \mathbb{P}\left(c(W_n^{L[b_1]}, e_1'e_1) = k\right)$ la probabilité d'avoir k structures en Y dans une plante d'âge n. Tout comme pour l'exemple précédent, la distribution $\left\{p_N^k\right\}_{k\geq 0}$ peut être déterminée récursivement en identifiant les coefficients des fonctions génératrices développées en séries entières :

$$\begin{aligned} \forall n \in \{0,\dots,N-1\}, \quad \forall k \geq 0, \\ p_{n+1}^k =& \delta_0(k)\left[1 - p_s(1) + p_s(1)(1-p_s(1))\right] + 2p_s(1)(1-p_s(1))p_n^k \\ &+ \mathbb{1}_{\mathbb{N}^*}(k)p_s(1)\sum_{q=0}^{k-1}\left(p_n^q + \delta_0(q)\left[p_s(1)-1\right]\right)\left(p_n^{k-1-q} + \delta_{k-1}(q)\left[p_s(1)-1\right]\right). \end{aligned}$$

Il est également possible d'obtenir récursivement les moments à tout ordre du nombre de fruits dans une plante d'âge N à partir de l'équation (5.7) grâce au corollaire A.1.2 de l'annexe A.1.

5.2.4 Distribution du nombre d'apex

Soit le modèle de développement stochastique $\mathcal{S}$ défini de la façon suivante :
- La classe physiologique maximale est $CP_m = 1$;
- Un entrenœud de CP 1 porte deux bourgeons axillaires de CP 1, aucun bourgeon apical et une feuille ;

– Un bourgeon de CP 1 est actif et produit un nouveau phytomère avec une probabilité $p_a(1)$ ou bien il est dormant avec une probabilité $1-p_a(1)$.

Les règles de production de $\mathcal{S}$ sont illustrées par la figure 5.11.

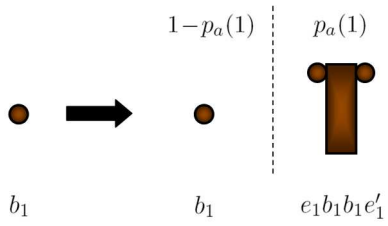

FIG. 5.11 – Règles de production avec dormance.

Nous souhaitons déterminer la distribution associée au nombre d'apex dans une plante générée aléatoirement par $\mathcal{S}$ après N cycles de développement. Un apex est en fait le bout d'une tige, cf figure 5.12.

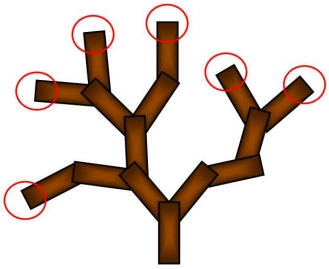

FIG. 5.12 – Exemples d'apex.

Etant donné que nous ne nous intéressons pas aux feuilles et que celles-ci n'interviennent pas dans la mise en place de l'architecture de la plante, nous les omettons une fois de plus pour la suite de l'étude. Les ensembles de symboles associés aux bourgeons et aux autres organes de la plante sont alors donnés respectivement par $\mathcal{B} = \{b_1\}$ et $\mathcal{O} = \{e_1\}$. En tenant compte des conventions de notations pour les mots de Dyck associés aux structures de plante (voir section 5.1.2), l'alphabet est donc donné par :

$$V = \{b_1, e_1, e'_1\}.$$

La matrice de transition π associée aux règles de production de $\mathcal{S}$ est la matrice de V dans W dont toutes les composantes sont nulles à l'exception de :

$$\pi_{b_1,b_1} = 1-p_a(1) \qquad \pi_{b_1,e_1b_1b_1e'_1} = p_a(1) \qquad \pi_{e_1,e_1} = 1 \qquad \pi_{e'_1,e'_1} = 1$$

Soit $L = \langle W^+, \pi \rangle$ le F0L-système stochastique associé au modèle de développement $\mathcal{S}$. Déterminer la distribution associée au nombre d'apex dans une plante générée aléatoirement par $\mathcal{S}$ après N cycles de développement est équivalent à déterminer la distribution associée au nombre d'occurrences du mot $e_1b_1b_1e'_1$ dans un texte généré aléatoirement par le système composant $L[b_1]$ après N étapes de production. Pour résoudre le problème,

nous souhaitons calculer $\Psi_N^{L[b_1]}$, la fonction génératrice de $L[b_1]$ associée à $\{e_1b_1b_1e_1'\}$ après N étapes de production :

$$\Psi_N^{L[b_1]}(z) = \sum_{w\in W} (P^N)_{b_1,w}\, z^{c(w,e_1b_1b_1e_1')} = \sum_{k\in\mathbb{N}} \mathbb{P}\left(c(W_N^{L[b_1]}, e_1b_1b_1e_1') = k\right) z^k$$

avec P le noyau de Markov associé à L et $\left(W_n^{L[b_1]}\right)_{n\geq 0}$ la chaîne de Markov engendrée par $L[b_1]$. $\Psi_N^{L[b_1]}(z)$ est encore déterminée récursivement en utilisant la méthode symbolique du chapitre 4. La première étape consiste à trouver une spécification itérative adéquate pour $\{W_n^{L[b_1]} \mid n \in \{0,\dots,N\}\}$. Utilisons tout d'abord le théorème de décomposition générale 4.3.2 et le théorème d'indépendance d'évolution 4.3.3 :

$$\forall n \in \{0,\dots,N-1\}, \qquad W_{n+1}^{L[b_1]} = \{(\epsilon, 1-p_s(1))\}.W_n^{L[b_1]} + \{(e_1, 1-p_s(1))\}.W_n^{L[b_1]}.W_n^{L[b_1]}.\{(e_1', 1)\}$$

Le système d'équations combinatoires précédent n'est pas une bonne spécification car la concaténation de $\{(e_1, 1-p_s(1))\}$, $W_n^{L[b_1]}$, $W_n^{L[b_1]}$ et $\{(e_1', 1)\}$ est génératrice par rapport au mot $e_1b_1b_1e_1'$. L'idée est alors d'introduire la classe combinatoire stochastique $W_n^{L[e_1b_1b_1e_1']}$ dans la décomposition de $W_{n+1}^{L[b_1]}$. A partir de l'équation précédente, nous avons alors :

$$\forall n \in \{0,\dots,N-1\}, \qquad W_{n+1}^{L[b_1]} = \{(\epsilon, 1-p_s(1))\}.W_n^{L[b_1]} + \{(\epsilon, 1-p_s(1))\}.W_n^{L[e_1b_1b_1e_1']} \tag{5.8}$$

Cette fois, toutes les concaténations sont non génératrices par rapport à $e_1b_1b_1e_1'$. Pour que la spécification soit complète, il faut encore un système d'équations combinatoires exprimant $W_n^{L[b_1]}$ récursivement et reposant sur des constructions admissibles. Nous l'obtenons à partir de l'équation (5.8) :

$$\begin{aligned}
&\forall n \in \{0,\dots,N-1\},\\
&W_{n+1}^{L[e_1b_1b_1e_1']} = W_{n+1}^{L[e_1]}.W_{n+1}^{L[b_1]}.W_{n+1}^{L[b_1]}.W_{n+1}^{L[e_1']} = \{(e_1,1)\}.W_{n+1}^{L[b_1]}.W_{n+1}^{L[b_1]}.\{(e_1',1)\}\\
&\quad = \{(\epsilon,(1-p_s(1))^2)\}.W_n^{L[e_1b_1b_1e_1']} + \{(e_1, p_s(1)(1-p_s(1)))\}.W_n^{L[e_1b_1b_1e_1']}.W_n^{L[b_1]}.\{(e_1',1)\}\\
&\qquad + \{(e_1, p_s(1)(1-p_s(1)))\}.W_n^{L[b_1]}.W_n^{L[e_1b_1b_1e_1']}.\{(e_1',1)\}\\
&\qquad + \{(e_1, p_s(1)^2)\}.W_n^{L[e_1b_1b_1e_1']}.W_n^{L[e_1b_1b_1e_1']}.\{(e_1',1)\}
\end{aligned} \tag{5.9}$$

Les équations (5.8) et (5.9) forment bien une spécification itérative adéquate pour $\{W_n^{L[b_1]} \mid n \in \{0,\dots,N\}\}$. En utilisant les règles de transformation du théorème 4.3.1, nous obtenons le système d'équations fonctionnelles suivant :

$$\begin{aligned}
&\forall n \in \{0,\dots,N-1\}\\
&\begin{cases}
\Psi_{n+1}^{L[b_1]}(z) = (1-p_a(1))\Psi_n^{L[b_1]}(z) + p_a(1)\Psi_n^{L[e_1b_1b_1e_1']}(z)\\
\Psi_{n+1}^{L[e_1b_1b_1e_1']}(z) = (1-p_a(1))^2\Psi_n^{L[e_1b_1b_1e_1']}(z) + 2p_a(1)(1-p_a(1))\Psi_n^{L[e_1b_1b_1e_1']}(z)\Psi_n^{L[b_1]}(z)\\
\qquad\qquad +p_a(1)^2 z\left(\Psi_n^{L[e_1b_1b_1e_1']}(z)\right)^2
\end{cases}
\end{aligned} \tag{5.10}$$

avec $\Psi_0^{L[b_1]}(z) = 1$ et $\Psi_0^{L[e_1b_1b_1e_1']}(z) = z$. On peut extraire des équations précédentes la probabilité $p_n^k = \mathbb{P}\left(c(W_n^{L[b_1]}, e_1b_1b_1e_1') = k\right)$ d'avoir k apex dans une plante d'âge n. On introduit également $q_n^k = \mathbb{P}\left(c(W_n^{L[e_1b_1b_1e_1']}, e_1b_1b_1e_1') = k\right)$ la probabilité d'avoir k apex dans une plante d'âge n. On en déduit :

$$\forall n \in \{0, \ldots, N-1\}, \quad \forall k \geq 0,$$
$$\begin{cases} p_{n+1}^k = (1 - p_a(1))p_n^k + p_a(1)q_n^k \\ q_{n+1}^k = (1 - p_a(1))^2 q_n^k + 2p_a(1)(1 - p_a(1)) \sum_{l=0}^{k} q_n^l p_n^{k-l} + \mathbb{1}_{\mathbb{N}^*}(k) \sum_{l=0}^{k-1} q_n^l p_n^{k-1-l} \end{cases}$$

En utilisant le corollaire A.1.2 de l'annexe A.1 et le système d'équations (5.10), il est également possible d'en déduire les moments à tout ordre du nombre d'apex dans une plante d'âge N.

5.3 Estimation des paramètres du modèle de développement

Nous proposons dans cette section une méthode permettant d'estimer le vecteur de paramètres Θ_{dev} associé au modèle $\mathcal{S}$ à partir de données botaniques (voir Loi et al. (2010)). Θ_{dev} est en fait le vecteur formé à partir des vecteurs Θ_{org} et Θ_{dif} (cf le chapitre 3) associés respectivement au modèle d'organogenèse et de différenciation de $\mathcal{S}$. Pour estimer Θ_{dev}, nous proposons une méthode s'inspirant de celle utilisée pour estimer le vecteur de paramètres Θ_{org}^{GL} associé au modèle d'organogenèse de GreenLab 2 (voir Wang et al. (2009) entre autres et la section 3.1.2 pour une description du modèle) mais qui combine également les résultats obtenus dans la section 5.2 grâce à la méthode symbolique. Nous présentons d'abord la méthode d'estimation des paramètres de l'organogenèse de GreenLab 2, cf section 5.3.1. Nous proposons ensuite une méthode pour l'estimation de Θ_{dev} associé au modèle de développement $\mathcal{S}$, section 5.3.2. Enfin, nous terminons par la présentation d'une procédure d'échantillonnage permettant de recueillir les données à partir d'une population de plantes (section 5.3.3).

5.3.1 Cas du modèle GreenLab 2

Dans GreenLab 2, l'estimation de Θ_{org}^{GL} se fait à partir de données botaniques liées à la composition de la plante comme par exemple le nombre de phytomères dans une structure donnée ou encore le nombre d'unités de croissance (*i.e.* l'ensemble des phytomères issus d'un bourgeon au cours d'une étape d'organogenèse). Ces données sont recueillies à partir d'une population de plantes dont l'âge maximal (en cycles de développement) est noté N_{max}. Pour la suite, nous choisissons de travailler avec le nombre de phytomères. Le principe reste le même pour tout autre type d'organe. Pour toute la suite de la section, nous faisons l'hypothèse de connaître l'âge de tous les phytomères des différentes plantes. L'estimation $\hat{\Theta}_{org}^{GL}$ de Θ_{org}^{GL} est alors choisie comme le vecteur minimisant la distance euclidienne entre deux vecteurs $\mathcal{M}_{emp}$ et $\mathcal{M}_{th}$ de $\mathbb{R}^M$, $M \in \mathbb{N}^*$, introduits de la façon suivante :

- construction de $\mathcal{M}_{emp}$: à partir des plantes, nous en déduisons l'ensemble des distributions empiriques $D_{emp}^{i,k}$ associées au nombre de phytomères présents dans une structure de CP $i \in \{1, \ldots, CP_m\}$ et d'âge $k \in \{1, \ldots, N_{max}\}$ (cf la section 5.3.3 pour l'obtention de $D_{emp}^{i,k}$). Pour tous les couples (i, k), il est alors possible de calculer la moyenne empirique $M_{emp}^{i,k}$ et la variance empirique $V_{emp}^{i,k}$ associées à $D_{emp}^{i,k}$. Le vecteur $\mathcal{M}_{emp}$ regroupe alors toutes les moyennes et variances empiriques possibles obtenues à partir des vraies plantes :

$$\mathcal{M}_{emp} = (M_{emp}^{i,k}, V_{emp}^{i,k})_{i \in \{1,\ldots,CP_m\}, k \in \{1,\ldots,N_{max}\}}.$$

- construction de $\mathcal{M}_{th}$: en utilisant la théorie des processus composés, il est possible de donner une expression explicite de $M_{th}^{i,k}$ et de $V_{th}^{i,k}$ qui sont respectivement l'espérance et la variance théorique associées au nombre de phytomères dans une structure de CP $i \in \{1, \ldots, CP_m\}$ et d'âge $k \in \{1, \ldots, N_{max}\}$ générée selon le modèle GreenLab 2 (voir Kang et al. (2008)). Le vecteur $\mathcal{M}_{th}$ regroupe alors l'ensemble de ces espérances et variances :

$$\mathcal{M}_{th} = (M_{th}^{i,k}, V_{th}^{i,k})_{i \in \{1,\ldots,CP_m\}, k \in \{1,\ldots,N_{max}\}}.$$

  et dépend du vecteur de paramètres Θ_{org}^{GL}.

L'estimation $\hat{\Theta}_{org}^{GL}$ est alors donnée par :

$$\hat{\Theta}_{org}^{GL} = \underset{\Theta_{org}^{GL} \in [0;1]^{dim(\Theta_{org}^{GL})}}{\operatorname{argmin}} ||\mathcal{M}_{th}(\Theta_{org}^{GL}) - \mathcal{M}_{emp}||_2^2 \tag{5.11}$$

avec $||.||_2$ la norme euclidienne sur $\mathbb{R}^{dim(\Theta_{org}^{GL})}$. Des algorithmes d'optimisation de type Levenberg-Marquardt (voir l'annexe B) peuvent être utilisés pour calculer $\hat{\Theta}_{org}^{GL}$.

5.3.2 Procédure utilisant la méthode symbolique

La méthode de GreenLab 2 peut également être utilisée pour fournir une estimation $\hat{\Theta}_{dev}$ du vecteur de paramètres Θ_{dev} associé au modèle $\mathcal{S}$. Cependant, il n'est pas toujours facile de compter le nombre de phytomères (ou le nombre d'unités de croissance) dans une plante. En effet, il est possible que les cicatrices laissées par les étapes successives d'organogenèse s'estompent avec le temps rendant difficile le comptage. Ceci entraîne alors une sous ou une sur-estimation des moyennes et variances empiriques $M_{emp}^{i,k}$ et $V_{emp}^{i,k}$ ce qui a pour effet de biaiser l'estimation $\hat{\Theta}_{dev}$ de Θ_{dev}. Au lieu de travailler avec des organes, nous proposons de travailler avec des motifs qui sont plus facilement identifiables à l'échelle de la plante (apex, structure en Y, ...). En effet, la méthode symbolique appliquée aux plantes permet de calculer récursivement moyennes et variances théoriques associées au nombre d'occurrences d'un motif particulier dans une structure donnée (voir les sections 5.2.3 et 5.2.4). Pour la suite, choisissons comme motif les structures en Y et supposons que l'on dispose toujours d'une population de plantes dont l'âge maximal (en cycles de développement) est noté N_{max} et que nous sommes capables de donner l'âge de

chacune des plantes. Notre méthode propose une estimation $\hat{\Theta}_{dev}$ de Θ_{dev} choisie comme le vecteur minimisant la distance euclidienne entre deux vecteurs $\mathcal{M}_{emp}$ et $\mathcal{M}_{th}$ de $\mathbb{R}^M$, $M \in \mathbb{N}^*$, introduits de la façon suivante :

- construction de $\mathcal{M}_{emp}$: à partir des vraies plantes, nous en déduisons l'ensemble des distributions empiriques D^k_{emp} associées au nombre de structures en Y présentes dans une plante d'âge $k \in \{1, \ldots, N_{max}\}$ (cf la section 5.3.3). Il est donc possible de calculer la moyenne empirique M^k_{emp} et la variance empirique V^k_{emp} associées à D^k_{emp}. Le vecteur $\mathcal{M}_{emp}$ regroupe alors toutes les moyennes et variances empiriques possibles obtenues à partir des vraies plantes :

$$\mathcal{M}_{emp} = (M^k_{emp}, V^k_{emp})_{k \in \{1,\ldots,N_{max}\}}.$$

- construction de $\mathcal{M}_{th}$: en utilisant la méthode symbolique dans le cadre des plantes, il est possible de calculer récursivement M^k_{th} et V^k_{th} qui sont respectivement l'espérance et la variance théorique associées au nombre de structures en Y dans une plante d'âge $k \in \{1, \ldots, N_{max}\}$ générée selon le modèle $\mathcal{S}$. Le vecteur $\mathcal{M}_{th}$ regroupe alors l'ensemble de ces espérances et variances :

$$\mathcal{M}_{th} = (M^k_{th}, V^k_{th})_{k \in \{1,\ldots,N_{max}\}}.$$

  et dépend du vecteur de paramètres Θ_{dev}.

L'estimation $\hat{\Theta}_{dev}$ est alors donnée par :

$$\hat{\Theta}_{dev} = \underset{\Theta_{dev} \in D}{\operatorname{argmin}} \; ||\mathcal{M}_{th}(\Theta_{dev}) - \mathcal{M}_{emp}||^2_2 \tag{5.12}$$

avec $||.||_2$ la norme euclidienne sur $\mathbb{R}^{dim(\Theta_{dev})}$ et D un sous-ensemble borné de $\mathbb{R}^{dim(\Theta_{dev})}$. Tout comme pour GreenLab 2, des algorithmes d'optimisation de type Levenberg-Marquardt (voir l'annexe B) peuvent être utilisés pour résoudre le problème de minimisation. Dans le cas où le nombre de paramètres à estimer est supérieur au nombre d'équations disponibles, c'est-à-dire :

$$dim(\Theta_{dev}) > dim(\mathcal{M}_{emp}) = dim(\mathcal{M}_{th}),$$

il est toujours possible d'augmenter la dimension de $\mathcal{M}_{emp}$ (et donc celle de $\mathcal{M}_{th}$) en introduisant dans $\mathcal{M}_{emp}$ les moments d'ordre $l \geq 3$ concernant le nombre de structures en Y pour des plantes d'âge k et obtenus à partir de D^k_{emp}. Le vecteur $\mathcal{M}_{th}$ est alors lui aussi augmenté de façon similaire en introduisant les moments théoriques d'ordre $l \geq 3$ (toujours obtenus récursivement par la méthode symbolique).

Dans la méthode proposée, il est nécessaire de disposer d'un grand nombre de plantes (plusieurs centaines) pour construire des distributions empiriques $\{D^k_{emp}\}_{k \in \{1,\ldots,N_{max}\}}$ qui soient bien représentatives de la vraie distribution. En effet, un échantillon de petite taille engendrerait un biais important dans l'estimation du vecteur de paramètres Θ_{dev}. Si nous ne disposons seulement que de quelques plantes (de l'ordre de la dizaine), il n'est pas raisonnable d'appliquer la méthode précédente. Dans ce cas, il convient de

changer la nature des données botaniques utilisées. Au lieu de travailler à l'échelle de la plante entière, nous considérons cette fois l'ensemble des structures qui la compose. L'estimation $\hat{\Theta}_{dev}$ de Θ_{dev} est toujours choisie comme le vecteur minimisant la distance euclidienne entre deux nouveaux vecteurs $\mathcal{M}_{emp}$ et $\mathcal{M}_{th}$ de $\mathbb{R}^M$, $M \in \mathbb{N}^*$, introduits de la façon suivante :

- construction de $\mathcal{M}_{emp}$: à partir des vraies plantes, nous en déduisons l'ensemble des distributions empiriques $D^{i,k}_{emp}$ associées au nombre de structures en Y présentes dans une structure de CP $i \in \{1, \ldots, CP_m\}$ et d'âge $k \in \{1, \ldots, N_{max}\}$ (cf la section 5.3.3). De même, pour tous les couples (i, k), il est possible de calculer la moyenne empirique $M^{i,k}_{emp}$ et la variance empirique $V^{i,k}_{emp}$ associées à $D^{i,k}_{emp}$. Le vecteur $\mathcal{M}_{emp}$ regroupe alors toutes les moyennes et variances empiriques possibles obtenues à partir des vraies plantes :

$$\mathcal{M}_{emp} = (M^{i,k}_{emp}, V^{i,k}_{emp})_{i \in \{1,\ldots,CP_m\}, k \in \{1,\ldots,N_{max}\}}.$$

- construction de $\mathcal{M}_{th}$: en utilisant la méthode symbolique dans le cadre des plantes, il est possible de calculer récursivement $M^{i,k}_{th}$ et $V^{i,k}_{th}$ qui sont respectivement l'espérance et la variance théorique associées au nombre de structures en Y dans une structure de CP $i \in \{1, \ldots, CP_m\}$ et d'âge $k \in \{1, \ldots, N_{max}\}$ générée selon le modèle $\mathcal{S}$. Le vecteur $\mathcal{M}_{th}$ regroupe alors l'ensemble de ces espérances et variances :

$$\mathcal{M}_{th} = (M^{i,k}_{th}, V^{i,k}_{th})_{i \in \{1,\ldots,CP_m\}, k \in \{1,\ldots,N_{max}\}}.$$

  et dépend du vecteur de paramètres Θ_{dev}.

L'estimation $\hat{\Theta}_{dev}$ est encore donnée par :

$$\hat{\Theta}_{dev} = \underset{\Theta_{dev} \in D}{\operatorname{argmin}} \ ||\mathcal{M}_{th}(\Theta_{dev}) - \mathcal{M}_{emp}||^2_2 \tag{5.13}$$

avec $||.||_2$ la norme euclidienne sur $\mathbb{R}^{dim(\Theta_{dev})}$ et D un sous-ensemble borné de $\mathbb{R}^{dim(\Theta_{dev})}$. Cette nouvelle version de la méthode nécessite de connaître l'âge de tous les phytomères de la plante ce qui est une hypothèse très restrictive. En effet, dans le cas d'une plante avec développement stochastique, il se peut que les bourgeons de celle-ci marquent des pauses dans leur fonctionnement. En général, il n'est alors pas possible de donner l'âge des phytomères avec précision.

5.3.3 Procédure d'échantillonnage

Dans cette section, nous proposons une procédure concernant l'obtention des distributions empiriques D^k_{emp} et $D^{i,k}_{emp}$, $i \in \{1, \ldots, CP_m\}$ et $k \in \{1, \ldots, N_{max}\}$, intervenant dans les méthodes d'estimation des deux sections précédentes.

Obtention de D^k_{emp}, $k \in \{1, \ldots, N_{max}\}$:

Nous supposons toujours que nous disposons d'un ensemble de plantes dont l'âge maximal est N_{max} et que nous sommes capables de donner l'âge de chacune d'entre elles. La procédure est la suivante :

1. Choisir un organe ou un motif à partir duquel la méthode d'estimation se fera ;
2. Pour chaque plante, compter le nombre d'organes ou de motifs d'intérêt ;
3. A partir du modèle d'organogenèse, établir le nombre maximal O^k_{max} d'organes ou de motifs d'intérêt que peut contenir une plante d'âge k ;
4. Calculer la probabilité empirique $p^k(l)$ d'avoir l organes ou motifs dans une plante d'âge k :

$$p^k(l) = \frac{\text{nombre de plantes d'âge } k \text{ ayant } l \text{ organes ou motifs}}{\text{nombre total de plantes d'âge } k}$$

si le nombre de plantes d'âge k est différent de 0. Dans le cas contraire, on pose $p^k(l) = \delta_0(l)$.

5. On définit alors la distribution empirique D^k_{emp} comme le vecteur de taille $O^k_{max}+1$ tel que :

$$D^k_{emp} = (p^k(l))_{l\in\{0,\ldots,O^k_{max}\}}.$$

Obtention de $D^{i,k}_{emp}$, $i \in \{1,\ldots,CP_m\}$ et $k \in \{1,\ldots,N_{max}\}$:

Nous supposons toujours que nous disposons d'un ensemble de plantes dont l'âge maximal est N_{max} et que nous sommes capables de donner l'âge des différents phytomères de chacune d'entre elles. La procédure est la suivante :

1. Choisir un organe ou un motif à partir duquel la méthode d'estimation se fera ;
2. Isoler parmis les vraies plantes toutes les structures de CP i et d'âge k ;
3. Pour chacune de ces structures, compter le nombre d'organes ou de motifs d'intérêt ;
4. A partir du modèle d'organogenèse, établir le nombre maximal $O^{i,k}_{max}$ d'organes ou de motifs d'intérêt que peut contenir une structure de CP i et d'âge k ;
5. Calculer la probabilité empirique $p^{i,k}(l)$ d'avoir l organes ou motifs dans une structure de CP i et d'âge k :

$$p^{i,k}(l) = \frac{\text{nombre de structures de CP } i \text{ et d'âge } k \text{ ayant } l \text{ organes ou motifs}}{\text{nombre total de structures de CP } i \text{ et d'âge } k}$$

si le nombre de structures de CP i et d'âge k est différent de 0. Dans le cas contraire, on pose $p^{i,k}(l) = \delta_0(l)$.

6. On définit alors la distribution empirique $D^{i,k}_{emp}$ comme le vecteur de taille $O^{i,k}_{max}+1$ tel que :

$$D^{i,k}_{emp} = (p^{i,k}(l))_{l\in\{0,\ldots,O^{i,k}_{max}\}}.$$

Les figures 5.13 et 5.14 montrent deux procédures de comptage sur une plante pour l'obtention de $D^{i,k}_{emp}$, la première pour le nombre de phytomères et la seconde pour le nombre de structures en Y.

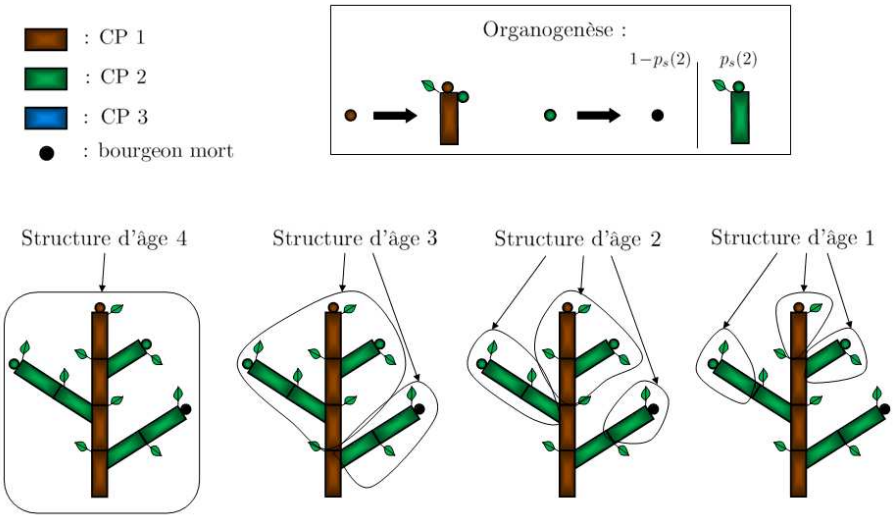

FIG. 5.13 – Comptage du nombre de phytomères sur une plante. Par exemple, il y a 3 structures d'âge 2. Une d'entre elles est de CP 1 et contient trois phytomères. Les deux autres sont de CP 2 et contiennent respectivement un et deux phytomères. La plante entière est d'âge 4 et contient neuf phytomères.

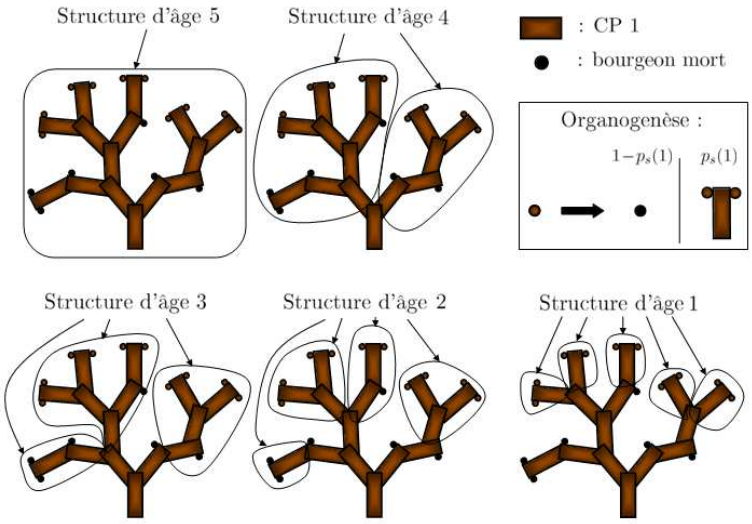

FIG. 5.14 – Comptage du nombre de structures en Y sur une plante. Par exemple, il y a deux structures d'âge 4 et de CP 1 contenant pour l'une une structure en Y et pour l'autre trois structures en Y. La plante entière est d'âge 5 et contient cinq structures en Y. On peut noter qu'il faut au moins deux cycles de développement pour créer une structure en Y. Les structures d'âge 1 ne contiennent donc aucune structure en Y.

Représentativité des distributions :

Les remarques suivantes concernent $\{D^{i,k}_{emp}\}_{i\in\{1,\dots,CP_m\},k\in\{1,\dots,N_{max}\}}$ mais sont également valables pour $\{D^{k}_{emp}\}_{k\in\{1,\dots,N_{max}\}}$.

Pour chaque couple $(i,k) \in \{1,\dots,CP_m\}\times\{1,\dots,N_{max}\}$, il est nécessaire de dispo-

ser d'un nombre suffisant de structures de CP i et d'âge k afin d'avoir une distribution empirique $D_{emp}^{i,k}$ représentative de la vraie distribution. Plus le nombre de structures disponibles est faible, plus le biais sur les paramètres estimés est important. Les tableaux 5.1 et 5.2 donnent une idée du biais sur des paramètres estimés pour une loi multinomiale en fonction de la taille de l'échantillon disponible. Un minimum de cent réalisations permet d'avoir une estimation des paramètres faiblement biaisée.

	Vraies valeurs	Est. $N = 30$	Est. $N = 100$	Est. $N = 1000$
p_1	$1/4 = 0.2500$	0.2517	0.2488	0.2495
p_2	$1/3 \approx 0.3333$	0.3299	0.3337	0.3335
p_3	$5/12 \approx 0.4167$	0.4184	0.4175	0.4170

TAB. 5.1 – Estimation des paramètres d'une multinomiale $\mathscr{M}(N, p_1, p_2, p_3)$. La procédure est la suivante : on affecte à N individus une caractéristique c_i avec une probabilité p_i, $i \in \{1, 2, 3\}$. L'estimation de la probabilité p_i correspond alors à la fréquence empirique associée à la caractéristique c_i dans la population. La valeur Est. (= Estimée) correspond à la moyenne des fréquences empiriques obtenues à partir de 1000 répétitions de l'expérience aléatoire.

	Vraies valeurs	Est. $N = 30$	Est. $N = 100$	Est. $N = 1000$
p_1	$1/4 = 0.2500$	0.2550	0.2510	0.2499
p_2	$1/6 \approx 0.1667$	0.1689	0.1684	0.1665
p_3	$1/5 \approx 0.2000$	0.1586	0.1989	0.2002
p_4	$1/3 \approx 0.3333$	0.2960	0.3343	0.3337
p_5	$1/20 \approx 0.0500$	0.1215	0.0505	0.0498

TAB. 5.2 – Estimation des paramètres d'une multinomiale $\mathscr{M}(N, p_1, p_2, p_3, p_4, p_5)$. La procédure est la même que celle du tableau 5.1

Si le nombre de structures de CP i et d'âge k est insuffisant, il est préférable de ne pas intégrer les moyennes et variances empiriques correspondantes $M_{emp}^{i,k}$ et $V_{emp}^{i,k}$ dans le vecteur $\mathcal{M}_{emp}$ afin de ne pas biaiser l'estimation de Θ_{dev}. Dans ce cas, il faut également retirer les composantes associées $M_{th}^{i,k}$ et $V_{th}^{i,k}$ dans $\mathcal{M}_{th}$. Il est alors possible que le nombre de paramètres à estimer donné par $dim(\Theta_{dev})$ soit supérieur au nombre d'équations disponibles donné par $dim(\mathcal{M}_{emp}) = dim(\mathcal{M}_{th})$. Dans ce cas là, on procède comme décrit à la fin de la section 5.3.2 en rajoutant des moments d'ordre $l \geq 3$ pour des distributions $D_{emp}^{i,k}$ bien représentatives (c'est-à-dire pour lesquelles nous avons suffisamment de données).

Procédure d'échantillonnage automatique :

La procédure de comptage peut s'avérer assez fastidieuse surtout pour des plantes dont la topologie est complexe. Une idée serait de pouvoir obtenir directement ces données à partir de techniques de digitalisation. De nombreux travaux ont été effectués sur la digitalisation de vraies plantes (voir Zhu et al. (2009) pour un ensemble de références

sur le sujet). Les techniques les plus populaires reposent sur l'utilisation de scanner laser 3D (voir Xu et al. (2005)) ou encore de photos digitales (voir Cheng et al. (2007) et Quan et al. (2006)). La figure 5.15 montre un exemple de digitalisation d'une vraie plante.

FIG. 5.15 – Digitalisation d'une vraie plante. Image extraite de Quan et al. (2006).

Une fois digitalisée, nous disposons d'un modèle 3D de la plante qui nous intéresse. En particulier, nous disposons d'un graphe représentant la topologie de la plante. A partir d'un algorithme de recherche sur le graphe, il est alors possible de compter le nombre d'organes ou de motifs d'intérêt dans la plante. Une limite de ce genre d'approche est la fiabilité de la reconstruction 3D de la plante. Lorsque celle-ci est recouverte de feuilles, certaines parties ne sont pas directement visibles. Des algorithmes de reconstruction permettent alors de corriger ce problème mais ne garantissent pas de reproduire la vraie topologie de la plante. Ceci entraîne donc des erreurs au niveau du graphe et donc au niveau du comptage des organes ou des motifs.

Chapitre 6

Inférence bayésienne pour les modèles de Markov cachés

Les modèles de Markov cachés constituent l'un des modèles statistiques les plus utilisés pour étudier des systèmes dynamiques évoluant à temps discret (voir Rabiner (1989) et Ephraim and Merhav (2002)). Ils se sont illustrés au travers de nombreuses applications : traitement de l'image (Bunke and Caelli (2001)), économétrie (Kim and Nelson (1999)), biologie (Koski (2001)), reconnaissance vocale (Jelinek (1997)) ... A une étape donnée, les systèmes dynamiques en question sont caractérisés par un état que l'on suppose caché (c'est-à-dire qu'il ne nous est pas possible de le mesurer directement à partir du système). Seules sont accessibles des observations qui sont liées à l'état caché. Les modèles probabilistes associés à l'évolution de l'état caché et à l'observation du système peuvent dépendre de paramètres qui sont inconnus *a priori*. L'objectif de ce chapitre est de présenter un ensemble de méthodes bayésiennes permettant l'estimation de ces paramètres à partir des observations du système (on parle alors d'estimation *a posteriori*).

Dans la section 6.2, nous présentons les modèles de Markov cachés ainsi qu'un cas particulier : les modèles à saut de Markov. Nous introduisons également un cadre statistique permettant l'estimation des paramètres du modèle grâce à l'utilisation de méthodes bayésiennes. La section suivante présente quatre méthodes bayésiennes permettant d'estimer l'état caché du système dynamique à un instant donné à partir des observations du système : le filtre de Kalman sans parfum, le filtre particulaire, le filtre particulaire convolé et le filtre particulaire Rao-Blackwellisé.

N.B. 6.1 A l'origine, le terme « système dynamique » est employé pour désigner un système évoluant de façon déterministe. De façon générale, nous utiliserons ce terme pour désigner tout système que son évolution soit déterministe ou stochastique.

6.1 Conventions de notation

Les conventions de notation suivantes sont adoptées pour toute la suite du chapitre. Considérons un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Alors :

- Les variables aléatoires sont notées en majuscule et leur réalisation en minuscule avec la même lettre. Par exemple, pour une variable aléatoire Z, une réalisation est notée z.
- On note d_Z la dimension d'un vecteur aléatoire Z.
- Si Z est une variable aléatoire à valeurs discrètes, alors nous noterons $\mathbb{P}(Z = z) \stackrel{\text{def}}{=} P(z)$.
- Si Z admet une densité de probabilité (par rapport à la mesure de Lebesgue), alors elle sera notée $p(z)$.
- Soient Y et Z deux variables aléatoires. Si la loi conditionnelle de Y sachant $Z = z$ admet une densité de probabilité alors elle sera noté $p(y|Z = z) \stackrel{\text{def}}{=} p(y|z)$. Si Y est à valeurs discrète, alors nous noterons $\mathbb{P}(Y = y|Z = z) \stackrel{\text{def}}{=} P(y|z)$.
- Concernant les algorithmes présentés dans ce chapitre, la notation $z \sim p(z)$ (resp. $z \sim P(z)$) signifie que l'on génère une valeur z pour la variable aléatoire Z selon la loi caractérisée par la densité $p(z)$ (resp. la fonction de probabilité $P(z)$).
- Soient $(Z_n)_{n\in\mathbb{N}}$ un processus aléatoire à temps discrets. Pour tout $(s, t) \in \mathbb{N} \times \mathbb{N}$, $s \leq t$, on note $Z_{s:t}$ le vecteur aléatoire suivant :

$$Z_{s:t} \stackrel{\text{def}}{=} (Z_s, Z_{s+1}, \ldots, Z_{t-1}, Z_t).$$

  Une réalisation de $Z_{s:t}$ sera notée $z_{s:t}$.
- On note $\mathcal{N}(m, \Sigma)$ la loi normale d'espérance m et de matrice de covariance Σ.
- On note $\mathcal{N}(x; m, \Sigma)$ la densité de la loi normale d'espérance m et de matrice de covariance Σ évaluée au point $x \in \mathbb{R}^d$ avec d la dimension de la loi normale considérée.

N.B. 6.2 Les conventions précédentes restent valables pour les vecteurs aléatoires.

6.2 Modèles statistiques

Dans cette section, nous présentons les différents modèles statistiques qui servent de cadre de travail pour l'estimation des paramètres du système dynamique d'intérêt. Nous introduisons d'abord les modèles de Markov cachés (cf section 6.2.1) et ensuite le cas particulier des modèles à saut de Markov (cf section 6.2.2). La section 6.2.3 montre comment adapter les modèles précédents pour permettre l'estimation des paramètres du système dynamique grâce à une approche bayésienne.

6.2.1 Modèle de Markov caché

Considérons un système dynamique évoluant à temps discret et dont l'état à un instant $k \in \mathbb{N}$ est caractérisé par un vecteur aléatoire X_k. Dans cette thèse, nous supposons que les vecteurs X_k, $k \in \mathbb{N}$, ont même dimension d_X et que ceux-ci prennent leurs valeurs dans un espace d'état $\mathcal{X} \subset \mathbb{R}^{d_X}$ muni de sa tribu Borélienne $B(\mathcal{X})$. Nous

supposons également que l'état du système à un instant k est caché, c'est-à-dire qu'il ne nous est pas possible de mesurer X_k directement à partir du système. Cependant, toujours à l'instant k, un certain nombre de mesures sont faites à partir du système et sont regroupées sous la forme d'un vecteur aléatoire Y_k. Les vecteurs Y_k sont à valeurs dans un espace $\mathcal{Y}_k \subset \mathbb{R}^{d_{Y_k}}$ muni de sa tribu Borélienne $B(\mathcal{Y}_k)$ et sont appelés observations du système. Enfin, pour tout $k \in \mathbb{N}$, nous supposons que X_k et Y_k admettent une densité par rapport à la mesure de Lebesgue notée respectivement $p(x_k)$ et $p(y_k)$.

N.B. 6.3 Dans cette thèse, les espaces $\mathcal{X}$ et $\mathcal{Y}_k$ avec $k \in \mathbb{N}$ sont euclidiens. Les définitions présentées par la suite peuvent cependant être étendues aux espaces polonais.

Les modèles de Markov cachés peuvent être présentés de la façon suivante (voir Doucet et al. (2001) et Cappé et al. (2005)) :

Définition 6.2.1 (Modèle de Markov caché) *Le processus à temps discret* $\{(X_k, Y_k)\}_{k\in\mathbb{N}}$ *est un modèle de Markov caché s'il est caractérisé par les équations d'état suivantes :*

$$\begin{cases} X_{n+1} = f_{n+1}(X_n, W_{n+1}, \Theta), & n \geq 0, \\ Y_n = g_n(X_n, V_n, \Theta), & n \geq 0, \end{cases} \tag{6.1}$$

avec :

- f_n *et* g_n *des fonctions boréliennes ;*
- $\{W_n\}_{n\geq 0}$ *et* $\{V_n\}_{n\geq 0}$ *des suites mutuellement indépendantes de vecteurs aléatoires i.i.d. ;*
- Θ *un vecteur de paramètres prenant ses valeurs dans un espace* $\mathcal{P} \subset \mathbb{R}^{d_\Theta}$.

N.B. 6.4 La définition précédente est en fait celle d'un modèle à espace d'état général (« general state space model » en anglais). Dans Cappé et al. (2005), les modèles de Markov cachés sont présentés comme un processus à temps discret $\{(X_k, Y_k)\}_{k\in\mathbb{N}}$ tel que :

- $\{X_k\}_{k\in\mathbb{N}}$ est une chaîne de Markov ;
- $\{Y_k\}_{k\in\mathbb{N}}$ est une suite de vecteurs aléatoires conditionnellement indépendants telle que la distribution conditionnelle de Y_k sachant $X_{0:k}$ et $Y_{0:k-1}$ ne dépend que de X_k.

Les modèles de Markov cachés ont donc d'abord été définis pour des processus $\{X_k\}_{k\in\mathbb{N}}$ à valeurs dans un espace d'état fini (Baum and Petrie (1966)). Cependant, le terme de « modèle de Markov caché » a depuis été également utilisé pour les modèles à espace d'état général. Par la suite, nous désignerons ces deux termes de façon équivalente par la défintion 6.2.1.

N.B. 6.5 Le vecteur Θ ne contient pas *a priori* les paramètres liés aux distributions des bruits W_n et V_n. Concernant le modèle $\mathcal{M}$, ces paramètres sont estimés par d'autres procédures (cf le chapitre 7).

Un modèle de Markov caché peut donc se décomposer en deux modèles :

- un **modèle d'évolution** donné par la première équation fonctionnelle du système (6.1) : ce modèle permet de construire l'état X_{n+1} du système à partir X_n. Le vecteur aléatoire W_n est appelé bruit du modèle ou bruit de modélisation et permet de prendre en compte des incertitudes au niveau du modèle ;
- un **modèle d'observation** donné par la seconde équation fonctionnelle du système (6.1) : ce modèle indique la façon dont les observations Y_n du système à l'instant n sont reliées à l'état du système X_n. Le vecteur aléatoire V_n est appelé bruit de mesure et représente la variabilité due au protocole de mesure.

Le vecteur de paramètres Θ est supposé inconnu *a priori*. Son estimation fait l'objet du présent chapitre. Nous pouvons noter que les fonctions f_n et g_n qui caractérisent respectivement le modèle d'évolution et le modèle de mesure dépendent de l'instant n considéré. Nous ne sommes donc pas dans un cadre homogène. Dans le cadre des modèles de Markov cachés, les dépendances entre vecteurs aléatoires peuvent être représentées à l'aide d'un diagramme de dépendance (également appelé modèle graphique de dépendance, voir Jensen (1996), Jordan (2004) et la figure 6.1).

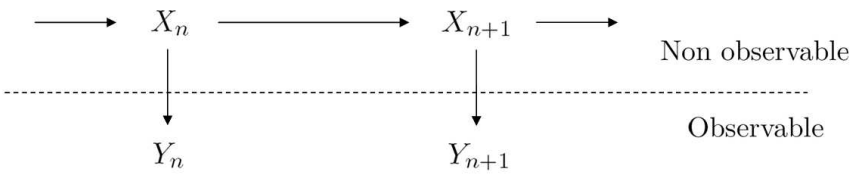

FIG. 6.1 – Diagramme de dépendance pour un modèle de Markov caché. Conditionnellement à $X_{0:n}$ et $Y_{0:n}$, la loi de X_{n+1} ne dépend que de X_n. De même, conditionnellement à $X_{0:n+1}$ et $Y_{0:n}$, la loi de Y_{n+1} ne dépend que de X_{n+1}

.

Un modèle de Markov caché peut donc être représenté de façon équivalente par le modèle statistique suivant :

$$\begin{cases} p(x_0), \\ p_\Theta(x_{n+1}|x_n), & n \geq 0, \\ p_\Theta(y_n|x_n), & n \geq 0, \end{cases} \tag{6.2}$$

avec $\Theta \in \mathcal{P}$.

6.2.2 Modèle à saut de Markov

Nous considérons toujours un système dynamique caractérisé par un processus à temps discrets $\{X_k, Y_k\}_{k\in\mathbb{N}}$ avec X_k l'état caché du système à l'instant k et Y_k son observation au même instant. Nous supposons également cette fois que le modèle d'évolution et le modèle d'observation à l'instant k dépendent d'un troisième vecteur aléatoire caché C_k et que tous les C_k ont même dimension d_C. Le processus à temps discret $\{C_k\}_{k\geq 0}$ évolue dans un espace d'état discret $\mathcal{C} \subset \mathbb{R}^{d_C}$ muni de la tribu formée de l'ensemble des sous-ensembles de $\mathcal{C}$. Les modèles à saut de Markov peuvent être présentés de la façon suivante (voir Doucet et al. (2001), Andrieu et al. (2003) et Cappé et al. (2005)) :

Définition 6.2.2 (Modèle à saut de Markov) *Supposons que $\{C_k\}_{k\geq 0}$ est une chaîne de Markov (généralement non-homogène) caractérisée par sa distribution initiale $P(c_0)$ et ses probabilités de transition $P(c_{n+1}|c_n)$. Le processus à temps discret $\{(X_k, Y_k)\}_{k\in\mathbb{N}}$ est un modèle à saut de Markov s'il est caractérisé par les équations d'état suivantes :*

$$\begin{cases} X_{n+1} = f_{n+1}(X_n, C_{n+1}, W_{n+1}, \Theta), & n \geq 0, \\ Y_n = g_n(X_n, C_n, V_n, \Theta), & n \geq 0, \end{cases} \tag{6.3}$$

avec :

- *f_n et g_n des fonctions boréliennes ;*
- *$\{W_n\}_{n\geq 0}$ et $\{V_n\}_{n\geq 0}$ des suites mutuellement indépendantes de vecteurs aléatoires i.i.d. ;*
- *Θ un vecteur de paramètres prenant ses valeurs dans un espace $\mathcal{P} \subset \mathbb{R}^{d_\Theta}$.*

Dans la littérature, le terme « modèle à saut de Markov » est en général réservé dans le cas où l'espace d'état de la chaîne de Markov $\mathcal{C}$ est fini. Si ce n'est pas le cas, le terme employé est « modèle de Markov caché hiérarchique » (hierarchical hidden Markov model en anglais, voir Fine et al. (1998) et Cappé et al. (2005)). Dans le cas du modèle de croissance de plantes M, l'espace $\mathcal{C}$ est infini mais discret (cf chapitre 7). Nous conserverons le terme de modèle à saut de Markov.

N.B. 6.6 Les modèles à saut de Markov sont en fait des cas particuliers des modèles de Markov cachés. En effet, si on considère le vecteur d'état (X_k, C_k) à l'instant k, alors le processus à temps discret $\{((X_k, C_k), Y_k)\}_{k\geq 0}$ est un modèle de Markov caché.

Dans un modèle à saut de Markov, les dépendances entre vecteurs aléatoires peuvent être représentées à l'aide du diagramme de dépendance de la figure 6.2.

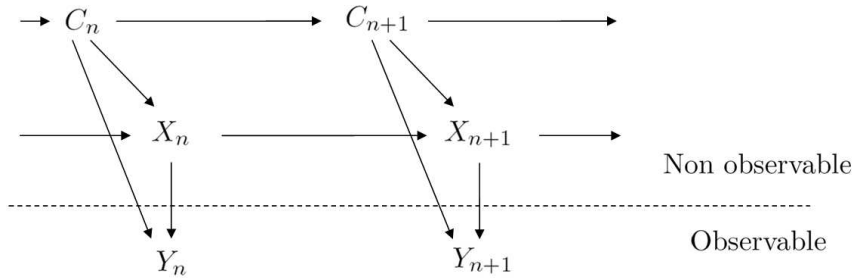

FIG. 6.2 – Diagramme de dépendance pour un modèle à saut de Markov. Conditionnellement au passé, la loi de C_{n+1} ne dépend que de C_n, la loi de X_{n+1} ne dépend que de X_n et de C_{n+1} et la loi de Y_{n+1} ne dépend que de X_{n+1} et de C_{n+1}.

Un modèle à saut de Markov peut donc être représenté de façon équivalente par le modèle statistique suivant :

$$\begin{cases} p(x_0), & \\ P(c_0), & \\ P(c_{n+1}|c_n), & n \geq 0, \\ p_\Theta(x_{n+1}|x_n, c_{n+1}), & n \geq 0, \\ p_\Theta(y_n|x_n, c_n), & n \geq 0, \end{cases} \tag{6.4}$$

avec $\Theta \in \mathcal{P}$. Le vecteur de paramètres Θ est toujours inconnu *a priori* et doit être estimé par les méthodes de la section 6.3.

6.2.3 Estimation bayésienne des paramètres du modèle

Les problématiques d'estimation associées aux modèles de Markov cachés et aux modèles à saut de Markov peuvent être classées dans trois catégories :

- le **filtrage** : pour $n \in \mathbb{N}$ donné, estimer l'état caché X_n à partir des observations $Y_{0:n}$. L'objectif est alors d'estimer la densité de probabilité conditionnelle $p(x_n|y_{0:n})$ (quand elle existe).
- le **lissage** : pour $(n, N) \in \mathbb{N}^2$ donné avec $n \leq N$, estimer l'état caché X_n à partir des observations $Y_{0:N}$. On cherche alors d'estimer la densité de probabilité conditionnelle $p(x_n|y_{0:N})$ (quand elle existe).
- la **prédiction** : pour $n \in \mathbb{N}$ donné, estimer l'état caché X_{n+1} à partir des observations $Y_{0:n}$. Il s'agit donc d'estimer la densité de probabilité conditionnelle $p(x_{n+1}|y_{0:n})$ (quand elle existe).

Lorsque l'on désire estimer l'état caché X_n et que les données arrivent au fur et à mesure, l'estimation est dite « en ligne ». Dans ce cas, il s'agit d'un problème de filtrage. Si, en revanche, nous disposons déjà d'un ensemble de données $Y_{0:N}$ et que l'on souhaite estimer l'état caché X_n avec $n \leq N$, alors l'estimation est dite « hors-ligne ». Il s'agit alors d'un problème de lissage. Les problèmes d'estimation hors-ligne peuvent également être traités par des méthodes de filtrage en chosissant d'ignorer $Y_{n+1:N}$ pour l'estimation de X_n. Les résultats obtenus sont généralement moins bons que pour les méthodes de lissage (ce qui se comprend puisque le lissage permet d'estimer avec plus d'informations). Dans le cadre de cette thèse, l'estimation est faite hors-ligne (voir la section sur les données du chapitre 7). Cependant, nous choisissons d'utiliser des méthodes de filtrage pour l'estimation des paramètres du modèle de croissance de plante. La raison est principalement due au rapport temps de calcul sur efficacité de l'estimation (voir la section 7.3.7 pour plus de détails). C'est pourquoi pour toute la suite de la thèse, nous ne nous intéresserons uniquement qu'aux méthodes de filtrage bien que l'estimation soit hors-ligne.

Les méthodes présentées dans la section 6.3 ont pour objectif d'estimer l'état caché X_n à partir d'un ensemble d'observations $Y_{0:n}$. Il est possible d'adapter le modèle statistique afin d'utiliser ces méthodes pour l'estimation du vecteur de paramètres Θ (cf les définitions 6.2.1 et 6.2.2). Il suffit en fait de l'intégrer dans le vecteur d'état caché du système dynamique en le considérant comme un vecteur aléatoire à valeurs dans l'espace mesurable $(\mathcal{P}, B(\mathcal{P}))$, voir Quach et al. (2007) et d'Alché Buc and Brunel (2010) pour des exemples en biologie moléculaire. Notons Θ_n le vecteur Θ associé à l'étape n. Soit X_n^a le vecteur d'état caché augmenté défini de la façon suivante :

$$X_n^a = (X_n, \Theta_n), \qquad n \geq 0.$$

L'espace d'état associé est $\mathcal{X}^a \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{P}$. Pour que le cadre statistique soit bien posé, il faut définir un modèle d'évolution pour le nouveau vecteur d'état X_n^a. Ceci implique en

fait de définir un modèle d'évolution pour Θ_n (c'est-à-dire une équation fonctionnelle permettant de construire Θ_{n+1} à partir de Θ_n). En règle générale, le vecteur Θ reste le même d'une étape à l'autre du système dynamique. Son évolution est donc constante *a priori*, c'est-à-dire $\Theta_n = \Theta_{n+1}$. Cependant, afin d'améliorer les performances de certaines méthodes bayésiennes (notamment celles du filtrage particulaire, voir le chapitre 7), nous pouvons attribuer à Θ_n une dynamique d'évolution artificielle. Nous supposerons alors que Θ_n évolue selon l'équation d'état :

$$\Theta_{n+1} = h_{n+1}(\Theta_n, U_{n+1}), \qquad n \geq 0,$$

avec h_n une fonction borélienne et $\{U_n\}_{n\geq 0}$ une suite de variables aléatoires i.i.d. de telle sorte que $\{U_n\}_{n\geq 0}$, $\{W_n\}_{n\geq 0}$ et $\{V_n\}_{n\geq 0}$ soient mutuellement indépendantes. Par la suite, pour tout $n \in \mathbb{N}$, nous supposons que Θ_n admet une densité de probabilité par rapport à la mesure de Lebesgue notée $p(\theta_n)$. En ce qui concerne les modèles de Markov cachés, nous pouvons donc définir une nouvelle fonction f_n^a de $\mathcal{X} \times \mathcal{P} \times \mathbb{R}^{d_{W_n}+d_{U_n}}$ dans $\mathcal{X} \times \mathcal{P}$ caractérisant le modèle d'évolution de X_n^a de la façon suivante :

$$\begin{aligned} \forall (x, \theta, (w, u)) \in \mathcal{X}\times\mathcal{P} \times \mathbb{R}^{d_{W_n}+d_{U_n}} \quad \text{et} \quad n \in \mathbb{N}, \\ f_n^a(x^a, w^a) = (f_n(x, w), h_n(\theta, u)) \end{aligned}$$

avec $x^a = (x, \theta)$ et $w^a = (w, u)$. f_n^a est définie de façon similaire pour les modèles à saut de Markov en intégrant simplement la dépendance en C_n dans la fonction.

Dans le cadre des modèles de Markov cachés, nous avons donc construit un nouveau système dynamique caractérisé par les équations d'état suivantes :

$$\begin{cases} X_{n+1}^a = f_{n+1}^a(X_n^a, W_{n+1}^a), & n \geq 0, \\ Y_n = g_n(X_n^a, V_n), & n \geq 0, \end{cases} \tag{6.5}$$

avec W_{n+1}^a la concaténation des vecteurs aléatoires W_{n+1} et U_{n+1}. Notons que nous avons toujours que $\{W_n^a\}_{n\geq 0}$ et $\{V_n\}_{n\geq 0}$ sont des suites mutuellement indépendantes de variables aléatoires i.i.d.. Dans le cas des modèles à saut de Markov, le nouveau système dynamique est le suivant :

$$\begin{cases} X_{n+1}^a = f_{n+1}^a(X_n^a, C_{n+1}, W_{n+1}^a), & n \geq 0, \\ Y_n = g_n(X_n^a, C_n, V_n), & n \geq 0, \end{cases} \tag{6.6}$$

avec W_{n+1}^a défini de façon similaire.

Les méthodes d'estimation de la section 6.3 permettent d'estimer le vecteur d'état caché X_n^a à partir des observations $Y_{0:n}$, estimation qui est notée $\hat{X}_n^a$. Il est alors possible d'en déduire une estimation $\hat{\Theta}_n$ de Θ en extrayant les composantes associées à Θ dans $\hat{X}_n^a$.

6.3 Quelques méthodes d'estimation bayésienne

Considérons un système dynamique évoluant à temps discret. Supposons que ce système peut être représenté par un modèle de Markov caché dont les équations d'états

sont données par (6.5) ou bien par un modèle de Markov à saut dont les équations d'états sont données par (6.6). Supposons également que nous disposons d'un ensemble d'observations $Y_{0:N}$ avec $N \in \mathbb{N}$. L'objectif de cette section est de présenter un ensemble de méthodes de filtrage bayésiennes permettant l'estimation de l'état caché X_n^a à partir des observations $Y_{0:n}$ avec $n \in \{0, \ldots, N\}$. Par la suite, nous notons $\hat{X}_n^a$ un estimateur de X_n^a. Nous supposons que la loi conditionnelle de X_n^a sachant $Y_{0:n}$ admet une densité de probabilité $p(x_n^a|y_{0:n})$ par rapport à la mesure de Lebesgue pour tout $n \in \{0, \ldots, N\}$. Le problème peut être résolu en construisant un estimateur $\hat{p}(x_n^a|y_{0:n})$ de $p(x_n^a|y_{0:n})$ et en choisissant l'un des deux estimateurs suivants pour $\hat{X}_n^a$ (voir Bar-Shalom et al. (2001)) :

- estimateur MMSE (= Minimum Mean Squared Error) :

$$\hat{X}_n^a = \mathbb{E}\left[X_n^a|Y_{0:n}\right] \approx \int_{\mathcal{X}^a} x_n^a \hat{p}(x_n^a|y_{0:n})\lambda(dx_n^a)$$

  avec λ la mesure de Lebesgue.

- estimateur MAP (= Maximum *a posteriori*) :

$$\hat{X}_n^a = \underset{x_n^a \in \mathcal{X}^a}{\operatorname{argmax}} \ \hat{p}(x_n^a|y_{0:n}).$$

Dans cette section, nous présentons quatre méthodes bayésiennes permettant de construire un estimateur $\hat{p}(x_n^a|y_{0:n})$ de $p(x_n^a|y_{0:n})$ avec $n \in \{0, \ldots, N\}$: filtre de Kalman sans parfum (Unscented Kalman Filter en anglais), filtre particulaire (Particle Filtering), filtre particulaire convolé (Convolution Particle Filter) et filtre particulaire de Rao-Blackwell (Rao-Blackwellized Particle Filter). Les trois premières méthodes peuvent être utilisées dans le cadre des modèles de Markov cachés et les trois dernières pour les modèles à saut de Markov. Dans chacune de ces méthodes, les densités $\hat{p}(x_n^a|y_{0:n})$, $n \in \{0, \ldots, N\}$, se calculent de façon récursive : l'étape n a pour but de déterminer $\hat{p}(x_{n+1}^a|y_{0:n+1})$ à partir d'un ensemble de relations entre $p(x_n^a|y_{0:n})$ et $p(x_{n+1}^a|y_{0:n+1})$ et se décompose en deux points :

- **Prédiction** : on détermine la densité de probabilité de prédiction de l'état du système :

$$p(x_{n+1}^a|y_{0:n}) = \int_{\mathcal{X}^a} p(x_{n+1}^a|x_n^a)p(x_n^a|y_{0:n})\lambda(dx_n^a). \tag{6.7}$$

- **Correction** : on corrige la densité de probabilité de prédiction en tenant compte de la nouvelle observation du système y_{n+1} pour obtenir $p(x_{n+1}^a|y_{0:n+1})$:

$$p(x_{n+1}^a|y_{0:n+1}) = \frac{p(x_{n+1}^a|y_{0:n})p(y_{n+1}|x_{n+1}^a)}{\int_{\mathcal{X}^a} p(x_{n+1}^a|y_{0:n})p(y_{n+1}|x_{n+1}^a)\lambda(dx_{n+1}^a)}. \tag{6.8}$$

Le détail des calculs menant aux équations précédentes est donné dans l'annexe C.1. On obtient ainsi un système d'équations reliant $p(x_n^a|y_{0:n})$ à $p(x_{n+1}^a|y_{0:n+1})$ pour $n \in \{0, \ldots, N-1\}$. A partir de la connaissance de $p(x_0^a)$, on peut donc déterminer ces densités de probabilité de filtrage. Dans la pratique, ces quantités sont rarement calculables de façon explicite. Elles le sont par exemple dans le cas d'un système dynamique

linéaire gaussien, cf le filtre de Kalman standard de la section 6.3.1. Concernant des systèmes dynamiques plus complexes, des approximations sont en général faites pour se ramener à des cas simples comme par exemple le filtre de Kalman sans parfum de la section 6.3.1. D'autres procédures proposent d'approcher numériquement ces intégrales par des méthodes stochastiques. Il s'agit entre autre des méthodes particulaires (voir les sections 6.3.2, 6.3.3 et 6.3.4). Ces méthodes connaissent un succès grandissant. Ceci est notamment dû au fait qu'elles peuvent s'appliquer à des modèles de structure très complexe sans avoir à effectuer des calculs de densité très difficiles. Dans sa thèse, Caron (2006) propose une bonne vue d'ensemble de la plupart des méthodes décrites par la suite.

6.3.1 Filtre de Kalman sans parfum

A l'origine, le filtre de Kalman a été conçu pour des systèmes dynamiques linéaires et gaussiens (voir Kalman (1960)). C'est le cas par exemple du système d'équations (6.5) si f_n^a et g_n sont des fonctions linéaires en X_n^a, en W_n^a et en V_n et si la densité de probabilité $p(x_0^a)$ associée à la loi de X_0^a est gaussienne. D'autres méthodes ont ensuite été développées dans le cas où f_n^a et g_n ne sont plus linéaires. Parmis elles, nous pouvons retenir le filtre de Kalman étendu (Extended Kalman Filter, Sorenson (1985)), le filtre de Kalman d'ensemble (Ensemble Kalman Filter, Evensen (1994)) et le filtre de Kalman sans parfum (Unscented Kalman Filter, Julier and Uhlmann (1997) et Quach et al. (2007)). Le filtre de Kalman étendu se ramène au filtre de Kalman standard en linéarisant les fonctions f_n^a et g_n par un développement de Taylor à l'ordre 1. Cependant, le filtre diverge rapidement si l'hypothèse de linéarité locale de f_n^a et g_n n'est pas respectée. Le filtre de Kalman d'ensemble permet d'estimer les matrices de covariance intervenant dans les équations du filtre standard (voir la section 6.3.1) par des méthodes de type Monte-Carlo. Cette méthode est assez coûteuse du point de vue du temps de calcul. Elle est très avantageuse dans le cas où la dimension du vecteur d'état est élevée (> 25, Luo and Moroz (2009)) ce qui n'est pas le cas pour cette thèse (voir le chapitre 7). Le filtre de Kalman sans parfum propose d'utiliser la transformation sans parfum (Unscented Transform, Julier et al. (2000) et Wan and Van Der Merwe (2000)) pour calculer les moments qui interviennent dans les équations du filtre. Cette méthode donne de bons résultats si les fonctions f_n^a et g_n ne sont pas fortement non-linéaires, auquel cas le filtre diverge. En raison des remarques précédentes, nous choisissons de travailler avec le filtre de Kalman sans parfum. Nous commençons tout d'abord par rappeler le principe du filtre de Kalman pour le cas linéaire gaussien et nous verrons comment l'adapter dans le cas non-linéaire avec la transformation sans parfum.

Dans toute cette section, nous supposons que W_n^a et V_n suivent des lois normales centrées de matrice de covariance respective J_n et R_n pour tout $n \in \mathbb{N}$.

Filtre de Kalman pour un système dynamique linéaire gaussien

Considérons le système dynamique défini par les équations (6.5) avec f_n^a une fonction linéaire en X_n^a et en W_n^a et g_n une fonction linéaire en X_n^a et additive en V_n. Supposons également que la densité de probabilité $p(x_0^a)$ associée à la loi de X_0^a est gaussienne. Dans ce cas, $p(x_n^a|y_{0:n})$ est la densité d'une loi normale d'espérance $\hat{x}_{n|n}^a$ et de matrice

de covariance $\hat{\Sigma}^{x^a}_{n|n}$ données par :

$$\begin{aligned}\hat{x}^a_{n|n} &= E[X^a_n|Y_{0:n}] \\ \hat{\Sigma}^{x^a}_{n|n} &= E[(X^a_n - \hat{x}^a_{n|n})^T(X^a_n - \hat{x}^a_{n|n})|Y_{0:n}].\end{aligned} \tag{6.9}$$

La connaissance de ces deux dernières quantités est alors suffisante pour caractériser $p(x^a_n|y_{0:n})$. Le filtre de Kalman permet de les déterminer séquentiellement en utilisant les équations (6.7) et (6.8) (voir Anderson and Moore (1979) et Arulampalam et al. (2002)). En effet, lorsque l'on traite des systèmes linéaires gaussiens, ces deux équations peuvent être calculées explicitement. L'étape n du filtre s'effectue en deux temps (on suppose que l'on a calculé $\hat{x}^a_{n|n}$ et $\hat{\Sigma}^{x^a}_{n|n}$) :

• **prédiction** (= étape de mise à jour) : l'objectif de cette étape est de déterminer $p(x^a_{n+1}|y_{0:n})$ (= densité de probabilité de prédiction de l'état du système) ainsi que $p(y_{n+1}|y_{0:n})$ (= densité de probabilité de prédiction de l'observation du système). Dans le cas d'un système linéaire gaussien, ces deux densités de probabilité sont celles de lois normales dont les espérances et matrices de covariance se déduisent des équations du système dynamique (6.5). L'espérance de X^a_{n+1} sachant $Y_{0:n}$ est alors donnée par :

$$\hat{x}^a_{n+1|n} = E[X^a_{n+1}|Y_{0:n}] = E[f^a_{n+1}(X^a_n, W^a_{n+1})|Y_{0:n}].$$

Comme le vecteur aléatoire W^a_{n+1} est indépendant de $\{W^a_k, V_{k'}\,;\,(k,k') \in \{1,\dots,n\}^2\}$, on en déduit que W^a_{n+1} est indépendant de $Y_{0:n}$. L'espérance intervenant dans le calcul de $\hat{x}^a_{n+1|n}$ est alors déterminée à partir des lois de probabilité $\mathcal{N}(0, J_{n+1})$ (= la loi de W^a_{n+1}) et de celle associée à $p(x^a_n|y_{0:n})$. On calcule de même la matrice de covariance associée à la loi de X^a_{n+1} sachant $Y_{0:n}$:

$$\hat{\Sigma}^{x^a}_{n+1|n} = E[(X^a_{n+1} - \hat{x}^a_{n+1|n})^T(X^a_{n+1} - \hat{x}^a_{n+1|n})|Y_{0:n}].$$

On procède de même pour $p(y_{n+1}|y_{0:n})$ et on en déduit l'espérance associée :

$$\hat{y}_{n+1|n} = E[Y_{n+1}|Y_{0:n}] = E[g_{n+1}(X^a_{n+1}) + V_{n+1}|Y_{0:n}] = E[g_{n+1}(X^a_{n+1})|Y_{0:n}] \tag{6.10}$$

ainsi que la matrice de covariance associée (en remarquant que V_{n+1} et $Y_{0:n}$ sont indépendants) :

$$\begin{aligned}\hat{\Sigma}^{y}_{n+1|n} &= E[(Y_{n+1} - \hat{y}_{n+1|n})^T(Y_{n+1} - \hat{y}_{n+1|n})|Y_{0:n}] \\ &= E[(g_{n+1}(X^a_{n+1}) - \hat{y}_{n+1|n})^T(g_{n+1}(X^a_{n+1}) - \hat{y}_{n+1|n})|Y_{0:n}] + R_{n+1}.\end{aligned} \tag{6.11}$$

La matrice de corrélation croisée entre X^a_{n+1} et Y_{n+1} conditionnellement à $Y_{0:n}$ intervient également dans les équations de l'étape de correction :

$$\hat{\Sigma}^{x^a y}_{n+1|n} = E[(X^a_{n+1} - \hat{x}^a_{n+1|n})^T(Y_{n+1} - \hat{y}_{n+1|n})|Y_{0:n}].$$

• **correction** : on corrige la densité de probabilité de prédiction de l'état du système $p(x^a_{n+1}|y_{0:n})$ en utilisant l'équation (6.8). L'espérance et la matrice de covariance de la loi normale associée à $p(x^a_{n+1}|y_{0:n+1})$ sont alors données par :

$$\begin{aligned}\hat{x}^a_{n+1|n+1} &= \hat{x}^a_{n+1|n} + K_{n+1}(y_{n+1} - \hat{y}_{n+1|n})^T \\ \hat{\Sigma}^{x^a}_{n+1|n+1} &= \hat{\Sigma}^{x^a}_{n+1|n} - K_{n+1}\hat{\Sigma}^{y}_{n+1|n}K^T_{n+1}.\end{aligned} \tag{6.12}$$

K_{n+1} s'appelle le gain de Kalman et vaut :

$$K_{n+1} = \hat{\Sigma}^{x^a y}_{n+1|n} \left(\hat{\Sigma}^{y}_{n+1|n} \right)^{-1}.$$

La quantité $y_{n+1} - \hat{y}_{n+1|n}$ s'appelle la correction. En regardant l'équation (6.12), on comprend que la prédiction de l'état du système $\hat{x}^a_{n+1|n}$ a été modifiée en fonction de la correction pour mieux correspondre à la nouvelle observation du système y_{n+1}. Plus la correction est importante, plus l'état prédit du système sera modifié.

Extension dans le cas non-linéaire

Lorsque les fonctions f^a_n et g_n ne sont plus linéaires en X^a_n, W^a_n (on conserve l'additivité de V_n), la densité $p(x^a_n|y_{0:n})$ n'est plus celle d'une loi gaussienne. Nous choisissons alors d'approcher cette densité par une autre densité $\hat{p}(x^a_n|y_{0:n})$ qui est celle d'une loi gaussienne dont l'espérance $\hat{x}^a_{n|n}$ et la matrice de covariance $\hat{\Sigma}^{x^a}_{n|n}$ sont toujours données par le système (6.9). Les quantités $\hat{x}^a_{n|n}$ et $\hat{\Sigma}^{x^a}_{n|n}$ sont alors calculées de façon récursive en utilisant les équations de mise-à-jour et de correction du filtre de Kalman standard. Cependant, à plusieurs reprises, nous sommes amenés à calculer l'espérance de la transformée par une fonction non-linéaire d'un vecteur aléatoire gaussien. Il n'est donc plus possible d'obtenir des relations de récurrence explicites comme dans le cas linéaire gaussien. Afin de contourner ce problème, nous utilisons une procédure de *transformation sans parfum*, cf annexe C.2. Prenons par exemple le calcul de $\hat{x}^a_{n+1|n}$ donné par :

$$\hat{x}^a_{n+1|n} = E[f^a_{n+1}(X^a_n, W^a_{n+1})|Y_{0:n}].$$

Etant donné que f^a_{n+1} n'est pas forcément linéaire en W^a_{n+1}, il est préférable de faire la transformée à partir du vecteur (X^a_n, W^a_{n+1}) au lieu de X^a_n. Ainsi, le calcul des sigma-points (cf l'annexe C.2) se fait à partir d'une distribution normale d'espérance :

$$\hat{x}^b_{n|n} = (\hat{x}^a_{n|n}, \mathbf{0}_{d_{W^a_{n+1}}})$$

et de matrice de covariance :

$$\hat{\Sigma}^{x^b}_{n|n} = \begin{pmatrix} \hat{\Sigma}^{x^a}_{n|n} & \mathbf{0}_{d_{x^a_n}, d_{W^a_{n+1}}} \\ \mathbf{0}_{d_{W^a_{n+1}}, d_{x^a_n}} & J_{n+1} \end{pmatrix}$$

avec $\mathbf{0}_k$ (resp. $\mathbf{0}_{k,k'}$) le vecteur de dimension k (resp. la matrice de taille k par k') dont toutes les composantes sont nulles. Notons $d_W = dim(W_{n+1})$, $d_X = dim(\hat{x}^a_{n|n})$ et $d_{W,X} = d_W + d_X$ et considérons les $2d_{W,X} + 1$ sigma-points $\chi^i_{n|n}$ associés au couple $\left(\hat{x}^b_{n|n}, \hat{\Sigma}^{x^b}_{n|n}\right)$ avec leur poids respectif ω_i pour $i \in \{1, \ldots, 2d_{W,X}+1\}$. Propageons chacun de ces sigma-points à travers l'équation d'évolution et nous obtenons $2d_{W,X}+1$ nouveaux sigma-points notés $\chi^i_{n+1|n}$ avec les mêmes poids :

$$\chi^i_{n+1|n} = f^a_{n+1}(\chi^i_{n|n}), \qquad i \in \{1, \ldots, M\}.$$

Dans ce cas, $\hat{x}^a_{n+1|n}$ est approchée par :

$$\hat{x}^a_{n+1|n} = \sum_{i=1}^{2d_{W,X}+1} \omega_i \, \chi^i_{n+1|n}.$$

N.B. 6.7 On suppose que l'état initial X_0^a suit une loi normale d'espérance x_0^a et de matrice de covariance $\Sigma_0^{x^a}$. x_0^a et $\Sigma_0^{x^a}$ sont supposées connues (ou fixées).

Le filtre de Kalman sans parfum est donc complètement décrit par l'algorithme suivant :

Algorithme 2 : Filtre de Kalman sans parfum

- **Initialisation**
- Poser $\hat{x}_{0|0}^a = x_0^a$ et $\hat{\Sigma}_{0|0}^{x^a} = \Sigma_0^{x^a}$.
- **Itération**
- Pour $n = 0, \ldots, N-1$:
  $ **Mise à jour** :
  - Poser $d_W = dim(W_{n+1})$, $d_X = dim(\hat{x}_{n|n}^a)$ et $d_{W,X} = d_W + d_X$
  - Poser :

$$\hat{x}_{n|n}^b = (\hat{x}_{n|n}^a, \mathbf{0}_{d_W}) \qquad \text{et} \qquad \hat{\Sigma}_{n|n}^{x^b} = \begin{pmatrix} \hat{\Sigma}_{n|n}^{x^a} & \mathbf{0}_{d_X,d_W} \cdot \\ \mathbf{0}_{d_W,d_X} & J_{n+1} \end{pmatrix}$$

  - Calculer les $2d_{W,X}+1$ sigma-points $\chi_{n|n}^i$ et leurs poids repectifs ω_i suivant une loi $\mathcal{N}\left(\hat{x}_{n|n}^b, \hat{\Sigma}_{n|n}^{x^b}\right)$.
  - Pour $i = 1, \ldots, 2d_{W,X}+1$, calculer le futur état augmenté du système donné par le i-ème sigma-point :

$$\chi_{n+1|n}^i = f_{n+1}^a(\chi_{n|n}^i).$$

  - Calculer l'estimation de l'état du système à l'instant $n+1$:

$$\hat{x}_{n+1|n}^a = \sum_{i=1}^{2d_{W,X}+1} \omega_i \chi_{n+1|n}^i$$

  ainsi que la matrice de covariance associée :

$$\hat{\Sigma}_{n+1|n}^{x^a} = \sum_{i=1}^{2d_{W,X}+1} \omega_i (\chi_{n+1|n}^i - \hat{x}_{n+1|n}^a)^T (\chi_{n+1|n}^i - \hat{x}_{n+1|n}^a).$$

  - Pour $i = 1, \ldots, 2d_{W,X}+1$, calculer la future observation du système donnée par le i-ème sigma-point :

$$\zeta_{n+1|n}^i = g_{n+1}(\chi_{n+1|n}^i).$$

  - Calculer l'estimation de l'observation du système à l'instant $n+1$:

$$\hat{y}_{n+1|n} = \sum_{i=1}^{2d_{W,X}+1} \omega_i \zeta_{n+1|n}^i$$

  ainsi que la matrice de covariance associée :

$$\hat{\Sigma}_{n+1|n}^{y} = \sum_{i=1}^{2d_{W,X}+1} \omega_i (\zeta_{n+1|n}^i - \hat{y}_{n+1|n})^T (\zeta_{n+1|n}^i - \hat{y}_{n+1|n}).$$

- Calculer la matrice de corrélation croisée :

$$\hat{\Sigma}^{x^a y}_{n+1|n} = \sum_{i=1}^{2d_{W,X}+1} \omega_i (\chi^i_{n+1|n} - \hat{x}^a_{n+1|n})^T (\zeta^i_{n+1|n} - \hat{y}_{n+1|n}).$$

$ **Correction** :
- Calculer le gain de Kalman :

$$K_{n+1} = \hat{\Sigma}^{x^a y}_{n+1|n} \left(\hat{\Sigma}^{y}_{n+1|n}\right)^{-1}.$$

- Calculer l'état corrigé du système à l'instant $n+1$:

$$\hat{x}^a_{n+1|n+1} = \hat{x}^a_{n+1|n} + K_{n+1}(y_{n+1} - \hat{y}_{n+1|n})^T$$

ainsi que sa matrice de covariance :

$$\hat{\Sigma}^{x^a}_{n+1|n+1} = \hat{\Sigma}^{x^a}_{n+1|n} - K_{n+1}\hat{\Sigma}^{y}_{n+1|n}K^T_{n+1}.$$

Pour des raisons de clarté, l'algorithme présenté ci-dessus ne prend pas en compte la première observation Y_0. Il est cependant possible de l'intégrer dans l'algorithme en rajoutant une étape dans laquelle les sigma-points sont créés à partir de la distribution $\mathcal{N}(x^a_0, \Sigma^{x^a}_0)$. Le reste de l'étape se déroule exactement de la même façon que pour n'importe quelle autre étape.

6.3.2 Filtre particulaire

Le filtrage particulaire est une méthode de type Monte-Carlo (Doucet et al. (2001), Doucet and Wang (2005) entre autres). L'objectif des méthodes de Monte-Carlo est de fournir une approximation numérique d'intégrales qui ne peuvent être calculées analytiquement. Cette approximation est obtenue grâce un échantillon de valeurs générées aléatoirement suivant une distribution d'intérêt. Les différents points abordés dans cette section sont issus des ouvrages Doucet et al. (2001) et Ristic et al. (2004) et des synthèses faites dans Caron (2006) et Davy (2006). Le filtrage particulaire s'applique aux modèles de Markov cachés mais également aux modèles à saut de Markov en considérant le vecteur d'état caché (X^a_n, C_n) à l'instant n (cf le *Nota Bene* 6.6).

De l'intégration de Monte-Carlo à l'échantillonnage d'importance

Soit Z un vecteur aléatoire à valeurs dans l'espace mesurable $\left(\mathbb{R}^{d_Z}, B(\mathbb{R}^{d_Z})\right)$ admettant une densité $p(z)$ par rapport à la mesure de Lebesgue. L'objectif est d'approcher numériquement l'espérance de $h(Z)$ avec h une fonction borélienne de $\left(\mathbb{R}^{d_Z}, B(\mathbb{R}^{d_Z})\right)$ dans $\left(\mathbb{R}^{l}, B(\mathbb{R}^{l})\right)$ avec $l \in \mathbb{N}^*$:

$$\mathbb{E}[h(Z)] = \int_{\mathbb{R}^{dZ}} h(z)p(z)\lambda(dz) \tag{6.13}$$

avec λ la mesure de Lebesque sur $\mathbb{R}^{d_Z}$. Lorsque la loi de Z est connue et facilement simulable, la méthode d'intégration de Monte-Carlo peut être utilisée pour résoudre ce problème (voir Pardoux (2007)). Considérons $\tilde{z}^{(i)}$, pour $i = 1, \ldots, M$, M réalisations de Z tirées selon la loi associée à la densité de probabilité $p(z)$. Dans ce cas, on peut faire l'approximation suivante :

$$\mathbb{E}[h(Z)] \approx \frac{1}{M} \sum_{i=1}^{M} h(\tilde{z}^{(i)}).$$

D'après la loi des grands nombres, moyennant quelques hypothèses sur h, il y a convergence presque sûr de la somme vers $\mathbb{E}[h(Z)]$ quand $M \to \infty$. Utiliser cette méthode revient en fait à approcher la loi de Z par la distribution suivante :

$$Z \sim \frac{1}{M} \sum_{i=1}^{M} \delta_{\tilde{z}^{(i)}}$$

avec $\delta_{\tilde{z}^{(i)}}$ la distribution de Dirac centrée en $\tilde{z}^{(i)}$. Cette méthode trouve cependant ses limites lorsque la loi de Z n'est pas connue ou difficile à simuler. La procédure d'échantillonnage d'importance a été introduite dans le but de contourner cette difficulté. L'idée est très proche de celle de l'intégration de Monte-Carlo et consiste à utiliser une autre loi caractérisée par une densité $\pi(z)$ (appelée densité d'importance ou *importance density* en anglais) pour la création de l'échantillon $\{\tilde{z}^{(i)},\ \ i = 1, \ldots, M\}$. L'intérêt de cette approche est de pouvoir choisir une loi facile à simuler. La densité $\pi(z)$ doit vérifier la condition :

$$\forall z \in \mathbb{R}^{d_Z},\ \ p(z) > 0 \quad \Rightarrow \quad \pi(z) > 0.$$

A partir de l'équation (6.13), nous obtenons :

$$\mathbb{E}[h(Z)] = \int_{\mathbb{R}^{d_Z}} h(z) \frac{p(z)}{\pi(z)} \pi(z) \lambda(dz) = \int_{\mathbb{R}^{d_Z}} h(z) w(z) \pi(z) \lambda(dz) \tag{6.14}$$

avec $w : z \mapsto p(z)/\pi(z)$. Suivant le même principe que l'intégration de Monte-Carlo classique, l'échantillonnage d'importance propose cette fois d'approcher la loi de Z (à une constante de normalisation près) comme suit :

$$Z \sim C. \sum_{i=1}^{M} w(\tilde{z}^{(i)}) \delta_{\tilde{z}^{(i)}}$$

avec $\tilde{z}^{(i)}$, pour $i = 1, \ldots, M$, M réalisations de Z tirées selon la loi caractérisée par la densité de probabilité $\pi(z)$. La constante de normalisation C est choisie de façon à ce que l'approximation de la loi de Z soit une loi de probabilité et vaut donc :

$$C = \left(\sum_{i=1}^{M} w(\tilde{z}^{(i)}) \right)^{-1}.$$

$w(\tilde{z}^{(i)})$ est appelé poids de la réalisation $\tilde{z}^{(i)}$ et $\tilde{w}(\tilde{z}^{(i)})$ son poids normalisé. L'approximation de $\mathbb{E}[h(Z)]$ est alors donnée par :

$$\mathbb{E}[h(Z)] \approx \sum_{i=1}^{M} \tilde{w}(\tilde{z}^{(i)}) h(\tilde{z}^{(i)}). \tag{6.15}$$

Dans le cadre des modèles de Markov cachés ou des modèles à saut de Markov, le calcul de l'estimateur MMSE $\hat{X}_n^a$ de X_n^a sachant $Y_{0:n}$ est donné par :

$$\hat{X}_n^a = \mathbb{E}[X_n^a|Y_{0:n}] = \int_{\mathcal{X}^a} x_n^a p(x_n^a|y_{0:n})\lambda(dx_n^a) = \int_{(\mathcal{X}^a)^{n+1}} h(x_{0:n}^a)p(x_{0:n}^a|y_{0:n})\lambda(dx_{0:n}^a).$$

avec $h : x_{0:n}^a \mapsto x_n^a$. Nous sommes donc bien dans la situation correspondant à cette section. $p(x_n^a|y_{0:n})$ est en fait une densité marginale de $p(x_{0:n}^a|y_{0:n})$. Sauf cas exceptionnels, la loi associée à la densité $p(x_{0:n}^a|y_{0:n})$ n'est pas connue ou est impossible à simuler. Nous devons alors utiliser l'échantillonnage d'importance pour résoudre le problème. La densité d'importance associée à $p(x_{0:n}^a|y_{0:n})$ est notée $\pi_n(x_{0:n}^a|y_{0:n})$. La fonction poids est notée w_n et est définie par :

$$\forall x_{0:n}^a \in (\mathcal{X}^a)^{n+1}, \quad w_n(x_{0:n}^a) = \frac{p(x_{0:n}^a|y_{0:n})}{\pi_n(x_{0:n}^a|y_{0:n})}.$$

Les poids normalisés $\tilde{w}_n$ sont définis comme précédemment. L'estimateur MMSE est alors obtenu par la relation suivante :

$$\hat{X}_n^a = \sum_{i=1}^{M} \tilde{w}(\tilde{x}_{0:n}^{a\,(i)})h(\tilde{x}_{0:n}^{a\,(i)}) = \sum_{i=1}^{M} \tilde{w}(\tilde{x}_{0:n}^{a\,(i)})\tilde{x}_n^{a\,(i)} \tag{6.16}$$

avec $\tilde{x}_{0:n}^{a\,(i)}$, pour $i = 1, \ldots, M$, M réalisations de $X_{0:n}^a$ sachant $Y_{0:n}$ tirées selon la loi caractérisée par la densité de probabilité $\pi_n(x_{0:n}^a|y_{0:n})$. Les réalisations $\{\tilde{x}_{0:n}^{a\,(i)}, \quad i = 1, \ldots, M\}$ sont appelées trajectoires. Intuitivement, une trajectoire $\tilde{x}_{0:n}^{a\,(i)}$ représente la propagation d'une particule caractérisée par un indice $i \in \{1, \ldots, M\}$ à travers le système dynamique pendant n étapes. Une particule i est en fait un point mobile de l'espace $\mathcal{X}^a$. La trajectoire $\tilde{x}_{0:n}^{a\,(i)}$ qui lui est associée représente les positions successives qu'elle a occupées entre les instants 0 et n. Le poids $w_n(\tilde{x}_{0:n}^{a\,(i)})$ associé à sa trajectoire est un indicateur de sa vraissemblance par rapport à la trajectoire réelle des états cachés du système dynamique.

Echantillonnage d'importance séquentiel

L'échantillonnage d'importance nécessite le calcul des poids $w_n(\tilde{x}_{0:n}^{(i)})$ pour $i = 1, \ldots, M$ ce qui revient à calculer de façon exacte $p(\tilde{x}_{0:n}^{a\,(i)}|y_{0:n})$ et $\pi_n(\tilde{x}_{0:n}^{a\,(i)}|y_{0:n})$. Cependant, ce calcul n'est pas toujours facile (voire possible) et peut s'avérer très fastidieux. Une façon de contourner le problème est d'utiliser le caractère markovien de l'évolution du système dynamique et de calculer les poids séquentiellement (c'est-à-dire en établissant une relation de récurrence entre $w_n(\tilde{x}_{0:n}^{a\,(i)})$ et $w_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)})$). Cette méthode porte le nom d'échantillonnage d'importance séquentiel ou encore de filtrage particulaire. Au lieu de raisonner directement sur la trajectoire entière ($= x_{0:n}^a$), on raisonne étape par étape (c'est-à-dire que l'on raisonne au niveau de la transition entre $x_{0:n}^a$ et x_{n+1}^a). On introduit tout d'abord la densité de transition d'importance $q_{n+1}(x_{n+1}^a|x_n^a)$ (*importance transition density* en anglais) permettant de propager les particules de l'étape n à l'étape $n+1$. Cette densité est choisie simple à simuler et de sorte à retranscrire la dynamique markovienne d'évolution du modèle de Markov caché associé au système dynamique (voir le

paragraphe concernant le choix de la densité d'importance avant l'algorithme du filtre particulaire). q_n doit également vérifier la condition suivante :

$$\forall x_{n+1}^a \in \mathcal{X}^a, \quad p(x_{n+1}^a|x_n^a) > 0 \quad \Rightarrow \quad q_n(x_{n+1}^a|x_n^a) > 0.$$

On définit alors la densité d'importance π_n par récurrence de la façon suivante :

$$\pi_{n+1}(x_{0:n+1}^a|y_{0:n+1}) \stackrel{\text{def}}{=} \pi_n(x_{0:n}^a|y_{0:n})q_{n+1}(x_{n+1}^a|x_n^a) = q_0(x_0^a)\prod_{k=0}^{n} q_{k+1}(x_{k+1}^a|x_k^a) \tag{6.17}$$

avec $q_0(x_0^a)$ la densité d'importance permettant de simuler X_0^a. Afin de calculer les poids $w_n(\tilde{x}_{0:n}^{a\,(i)})$ de façon séquentielle, on a recours à la formule de Bayes (voir le détail des calculs dans l'annexe C.3) :

$$p(x_{0:n+1}^a|y_{0:n+1}) = p(x_{0:n}^a|y_{0:n})\frac{p(y_{n+1}|x_{n+1}^a)p(x_{n+1}^a|x_n^a)}{p(y_{n+1}|y_{0:n})}. \tag{6.18}$$

En combinant les équations (6.17) et (6.18), nous obtenons alors :

$$w_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)}) = \frac{p(\tilde{x}_{0:n+1}^{a\,(i)}|y_{0:n+1})}{\pi_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)}|y_{0:n+1})} = w_n(\tilde{x}_{0:n}^{a\,(i)})\frac{p(y_{n+1}|\tilde{x}_{n+1}^{a\,(i)})p(\tilde{x}_{n+1}^{a\,(i)}|\tilde{x}_n^{a\,(i)})}{p(y_{n+1}|y_{0:n})q_{n+1}(\tilde{x}_{n+1}^{a\,(i)}|\tilde{x}_n^{a\,(i)})}. \tag{6.19}$$

Le calcul de $p(y_{n+1}|y_{0:n})$ est difficile mais pas nécessaire. En effet, il s'agit d'une constante de normalisation qui est la même pour toutes les trajectoires et qui disparaît donc dans le calcul des poids normalisés $\tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)})$:

$$\tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)}) = \frac{w_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)})}{\sum_{j=1}^{M} w_{n+1}(\tilde{x}_{0:n+1}^{a\,(j)})} \qquad \text{pour} \quad i = 1, \ldots, M. \tag{6.20}$$

En résumé, à la fin de l'étape n, on dispose d'un ensemble de M trajectoires $\tilde{x}_{0:n}^{a\,(i)}$ et des poids normalisés associés $\tilde{w}_n(\tilde{x}_{0:n}^{a\,(i)})$ pour $i = 0, \ldots, M$. On simule $\tilde{x}_{n+1}^{a\,(i)}$ selon la loi $q_{n+1}(\tilde{x}_{n+1}^{a\,(i)}|\tilde{x}_n^{a\,(i)})$ et on complète la trajectoire $\tilde{x}_{0:n+1}^{a\,(i)} \stackrel{\text{def}}{=} (\tilde{x}_{0:n}^{a\,(i)}, \tilde{x}_{n+1}^{a\,(i)})$ (il s'agit de l'étape de prédiction donnée par l'équation (6.7)). Les poids associés aux trajectoires $\tilde{x}_{0:n+1}^{a\,(i)}$ sont remis à jour par les équations (6.19) et (6.20) (cela correspond à l'étape de correction donnée par l'équation (6.8)).

Choix de la distribution d'importance

Le choix de la densité d'importance $q_n(x_n^a|x_{n-1}^a)$ (et donc de $\pi_n(x_{0:n}^a|y_{0:n})$) est primordiale pour s'assurer de l'efficacité du filtre particulaire. Cette densité gère la façon dont les particules vont parcourir l'espace des états $\mathcal{X}^a$ à l'instant n. Or, il se peut que cet espace soit très grand. Une densité d'importance permettant aux particules de parcourir l'espace sans guidage précis va entraîner une dégénerescence des poids des trajectoires. Ainsi, à un instant n, un grand nombre de trajectoires vont avoir un poids proche de zéro (ce qui correspond à des trajectoires improbables). Il en résulte une variance empirique au niveau des poids très importante ce qui nuit à la qualité des estimations telles que celle faite dans l'équation (6.16). La solution

consiste à parcourir intelligement l'espace des trajectoires. Supposons que l'on dispose d'une trajectoire à l'instant n, ${\tilde{x}^a_{0:n}}^{(i)}$, $i \in \{1, \ldots, M\}$. Pour orienter la particule dans la « bonne »direction, il faudrait tenir compte de la future observation du système y_{n+1} de sorte que la trajectoire complétée ${\tilde{x}^a_{0:n+1}}^{(i)}$ donne une observation proche de y_{n+1}. Dans cette optique, la densité d'importance serait $q_n(x^a_{n+1}|x^a_n) = p(x^a_{n+1}|x^a_n, y_{n+1})$. Il est prouvé que celle-ci est optimale au sens où c'est celle qui engendrerait une variance minimale des poids. Les particules se concentrent donc dans la « bonne »zone de l'espace des états. Cependant, d'après l'équation (6.19), il est nécessaire de pouvoir calculer $q_{n+1}({\tilde{x}^a_{n+1}}^{(i)}|\tilde{x}^{a(i)}_n) = p({\tilde{x}^a_{n+1}}^{(i)}|\tilde{x}^{a(i)}_n, y_{n+1})$ et ce calcul n'est généralement pas possible dans la majorité des systèmes dynamiques. Une solution plus simple est de choisir la densité d'importance $q_{n+1}(x^a_{n+1}|x^a_n) = p(x^a_{n+1}|x^a_n)$ correspondant à la probabilité de transition de la chaîne de Markov des états du système. Cette densité est plus facile à simuler et simplifie considérablement le calcul des poids mais elle entraîne une dégénerescence des poids plus importante. Elle reste cependant un bon compromis dans la plupart des cas. Il existe d'autres choix possibles pour la densité d'importance dont l'un consiste à prendre la densité d'évolution donnée par un pas de filtre de kalman (voir la section 6.3.1).

Rééchantillonnage

Malgré le choix d'une bonne densité d'importance, il est possible que les poids des particules dégénèrent. Dans ce cas, une stratégie consiste à choisir un nouveau jeu de particules en se concentrant sur les trajectoires à forte vraissemblance. Cette procédure s'appelle le rééchantillonnage. Lorsque la variance empirique des poids devient trop grande, on crée un nouveau jeu de particules en tirant aléatoirement avec remise M trajectoires parmi $\{{\tilde{x}^a_{0:n}}^{(i)}, i \in \{1, \ldots, M\}\}$ selon une multinomiale :

$$\mathscr{M}(1, \tilde{w}_n({\tilde{x}^a_{0:n}}^{(1)}), \tilde{w}_n({\tilde{x}^a_{0:n}}^{(2)}), \ldots, \tilde{w}_n({\tilde{x}^a_{0:n}}^{(M)}))$$

ce qui est cohérent puisque $\sum_{i=1}^{M} \tilde{w}_n({\tilde{x}^a_{0:n}}^{(i)}) = 1$. On affecte alors aux nouvelles trajectoires un poids $1/M$. En pratique (voir Kong et al. (1994) et Kitagawa (1996)), le rééchantillonnage est effectué lorsque :

$$\left[\sum_{i=1}^{M} (\tilde{w}_n({\tilde{x}^a_{0:n}}^{(i)}))^2\right]^{-1} \leq \eta$$

avec η un seuil de rééchantillonnage à définir (en général $\eta = 0.8M$, voir Caron (2006)). Cette condition repose sur le critère d'information (qui est lié à la variance empirique des poids des trajectoires). Il existe aussi d'autres procédures de rééchantillonage (voir par exemple Pham (2001)).

Algorithme

Algorithme 3 : Filtre Particulaire

• **Initialisation**
- Pour $i = 1, \ldots, M$, faire :
  - Générer $\tilde{x}_0^{a(1)} \sim q_0(x_0^a)$
  - Poser $w_0(\tilde{x}_0^{a(i)}) = \frac{p(y_0|\tilde{x}_0^{a(i)})p(\tilde{x}_0^{a(i)})}{q_0(\tilde{x}_0^{a(i)})}$

  \$ **Normalisation des poids** :
  - Pour $i = 1, \ldots, M$, faire :
    - Poser

$$\tilde{w}_0(\tilde{x}_0^{a(i)}) = w_0(\tilde{x}_0^{a(i)}) / \sum_{j=1}^{M} w_0(\tilde{x}_0^{a(j)})$$

  \$ **Rééchantillonnage** : si nécessaire, voir ci-dessous.

• **Itération**
- Pour $n = 0, \ldots, N-1$, faire :

  \$ **Prolongement des trajectoires** :
  - Pour $i = 1, \ldots, M$, faire :
    - Générer $\tilde{x}_{n+1}^{a}{}^{(i)} \sim q_{n+1}(x_{n+1}^a|\tilde{x}_n^{a(i)})$
    - Poser $\tilde{x}_{0:n+1}^{a}{}^{(i)} = (\tilde{x}_{0:n}^{a}{}^{(i)}, \tilde{x}_{n+1}^{a}{}^{(i)})$
    - Calculer

$$w_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(i)}) = \tilde{w}_n(\tilde{x}_{0:n}^{a}{}^{(i)}) \frac{p(y_{n+1}|\tilde{x}_{n+1}^{a}{}^{(i)})p(\tilde{x}_{n+1}^{a}{}^{(i)}|\tilde{x}_n^{a(i)})}{q_{n+1}(\tilde{x}_{n+1}^{a}{}^{(i)}|\tilde{x}_n^{a(i)})}$$

  \$ **Normalisation des poids** :
  - Pour $i = 1, \ldots, M$, faire :
    - Poser

$$\tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(i)}) = w_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(i)}) / \sum_{j=1}^{M} w_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(j)})$$

  \$ **Rééchantillonnage** :
  - Calculer $N_{\text{eff}} = \left[\sum_{i=1}^{M} (\tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(i)}))^2\right]^{-1}$
  - Si $N_{\text{eff}} \leq \eta$:
    - Pour $i = 1, \ldots, M$, faire :
      - Tirer avec remise une nouvelle trajectoire $\tilde{x}_{0:n+1}^{a}{}^{(i)}$ parmi $\{\tilde{x}_{0:n+1}^{a}{}^{(j)}, j \in \{1, \ldots, M\}\}$ selon une multinomiale $\mathscr{M}(1, \tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(1)}), \ldots, \tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(M)}))$
      - Poser $\tilde{w}_{n+1}(\tilde{x}_{0:n+1}^{a}{}^{(i)}) = 1/M$
  - Sinon, conserver le jeu de trajectoires avec leurs poids respectifs

---

6.3.3 Filtre particulaire convolé

Le filtre particulaire convolé est une méthode d'estimation non paramétrique de type filtrage particulaire (voir Campillo and Rossi (2009)). Tout comme le filtrage particulaire, cette méthode s'applique aux modèles de Markov cachés mais également aux modèles à saut de Markov en considérant le vecteur d'état caché (X_n^a, C_n) à l'instant n. Elle propose d'estimer $p(x_n^a|y_{0:n})$ de façon récurrente à l'aide d'un estimateur à noyau (voir l'annexe C.4) que l'on note $\hat{p}(x_n^a|y_{0:n})$. En tant que méthode bayésienne de filtrage,

elle se décompose aussi en deux étapes (on suppose que l'on dispose d'un estimateur $\hat{p}(x_n^a|y_{0:n})$) :

• **prédiction** : l'objectif est de fournir un estimateur à noyau de $p(x_{n+1}^a, y_{n+1}|y_{0:n})$ que l'on note $\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n})$. Pour cela, on a besoin, dans un premier temps, d'un ensemble E de réalisations indépendantes du couple (X_{n+1}^a, Y_{n+1}) conditionné par rapport à $Y_{0:n}$ à partir duquel on construit l'estimateur. Dans cette optique, on crée d'abord un jeu de M particules $\{\tilde{x}_n^{a\,(i)}, i = 1, \ldots, M\}$ tirées selon la loi de densité $\hat{p}(x_n^a|y_{0:n})$. On propage ensuite chacune de ces particules à travers l'équation d'évolution du système ce qui revient à obtenir un nouveau jeu de particules $\{\tilde{x}_{n+1^-}^{a\,(i)}, i = 1, \ldots, M\}$ tirées selon la loi de densité $p(x_{n+1}^a|\tilde{x}_n^{a\,(i)})$. Enfin, on propage chacune de ces dernières particules à travers l'équation d'observation du système ce qui revient à obtenir un nouveau jeu de particules $\{\tilde{y}_{n+1^-}^{(i)}, i = 1, \ldots, M\}$ tirées selon la loi de densité $p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})$. L'ensemble $\{(\tilde{x}_{n+1^-}^{a\,(i)}, \tilde{y}_{n+1^-}^{(i)}), i = 1, \ldots, M\}$ correspond à l'ensemble E des réalisations indépendantes recherché.

Dans un second temps, on construit l'estimateur à noyau $\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n})$ à partir de l'ensemble E. On choisit un noyau de Parzen-Rosenblatt pour le couple (X_{n+1}^a, Y_{n+1}) conditionné par rapport à $Y_{0:n}$ de la forme suivante :

$$K^{X,Y}_{h_M^X, h_M^Y}(x, y) = K^X_{h_M^X}(x) K^Y_{h_M^Y}(y)$$

avec $K^X_{h_M^X}$ et $K^Y_{h_M^Y}$ deux noyaux de Parzen-Rosenblatt et h_M^X (resp. h_M^Y) la taille de la fenêtre associée à $K^X_{h_M^X}$ (resp. $K^Y_{h_M^Y}$). On en déduit l'estimateur à noyau empirique de la densité de probabilité conditionnelle de (X_{n+1}^a, Y_{n+1}) sachant $Y_{0:n}$ (voir la définition C.4.3) :

$$\begin{aligned}\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n}) &= \frac{1}{M}\sum_{i=1}^{M} K^{X,Y}_{h_M^X, h_M^Y}(x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)}, y_{n+1} - \tilde{y}_{n+1^-}^{(i)}) \\ &= \frac{1}{M}\sum_{i=1}^{M} K^X_{h_M^X}(x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)}) K^Y_{h_M^Y}(y_{n+1} - \tilde{y}_{n+1^-}^{(i)}). \end{aligned} \tag{6.21}$$

• **correction** : dans cette étape, on propose un estimateur à noyau pour $p(x_{n+1}^a|y_{0:n+1})$. On l'obtient en utilisant l'équation (6.8) :

$$p(x_{n+1}^a|y_{1:n+1}) = \frac{p(x_{n+1}^a|y_{0:n})p(y_{n+1}|x_{n+1}^a)}{\int_{\mathcal{X}^a} p(x_{n+1}^a|y_{0:n})p(y_{n+1}|x_{n+1}^a)\lambda(dx_{n+1}^a)} = \frac{p(x_{n+1}^a, y_{n+1}|y_{0:n})}{\int_{\mathcal{X}^a} p(x_{n+1}^a, y_{n+1}|y_{0:n})\lambda(dx_{n+1}^a)}.$$

Etant donné que $p(x_{n+1}^a, y_{n+1}|y_{0:n})$ est estimée par $\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n})$, on obtient :

$$\begin{aligned}\int_{\mathcal{X}^a} p(x_{n+1}^a, y_{n+1}|y_{0:n})\lambda(dx_{n+1}^a) &\approx \frac{1}{M}\sum_{i=1}^{M}\left(\int_{\mathcal{X}^a} K^X_{h_M^X}(x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)})\lambda(dx_{n+1}^a)\right) K^Y_{h_M^Y}(y_{n+1} - \tilde{y}_{n+1^-}^{(i)}) \\ &\approx \frac{1}{M}\sum_{i=1}^{M} K^Y_{h_M^Y}(y_{n+1} - \tilde{y}_{n+1^-}^{(i)}).\end{aligned}$$

On obtient donc comme estimateur à noyau pour $p(x^a_{n+1}|y_{1:n+1})$:

$$\hat{p}(x^a_{n+1}|y_{1:n+1}) = \frac{\sum_{i=1}^{M} K^X_{h^X_M}(x^a_{n+1} - \tilde{x}^a_{n+1^-}{}^{(i)}) K^Y_{h^Y_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1^-})}{\sum_{i=1}^{M} K^Y_{h^Y_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1^-})}. \tag{6.22}$$

N.B. 6.8 La quantité $K^Y_{h^Y_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1^-}) / \sum_{i=1}^{M} K^Y_{h^Y_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1^-})$ peut s'interpréter comme le poids normalisé $\tilde{w}^{(i)}_{n+1}$ associé à la particule $\tilde{x}^a_{n+1^-}{}^{(i)}$.

N.B. 6.9 Dans le cas des modèles à saut de Markov, les équations de prédiction et de correction précédentes restent valables à condition de remplacer le noyau $K^X_{h^X_M}$ par le produit $K^X_{h^X_M} K^C_{h^C_M}$ avec $K^C_{h^C_M}$ le noyau de Parzen–Rosenblatt associé à C_n.

Algorithme 4 : Filtre Particulaire Convolé

- **Itération**
- Pour $n = -1, 0, \ldots, N-1$, faire :
  - $ **Prédiction** :
  - Pour $i = 1, \ldots, M$, faire :
    - Générer $\tilde{x}^a_{n+1^-}{}^{(i)} \sim p(x^a_{n+1}|\tilde{x}^a_n{}^{(i)})$
    - Générer $\tilde{y}^{(i)}_{n+1^-} \sim p(y_{n+1}|\tilde{x}^a_{n+1^-}{}^{(i)})$
    - Calculer les poids :
    $$w^{(i)}_{n+1} = K^Y_{h_M}(y_{n+1} - \tilde{y}^{(i)}_{n+1^-})$$
  - $ **Normalisation des poids** :
  - Pour $i = 1, \ldots, M$, faire :
    - Poser
    $$\tilde{w}^{(i)}_{n+1} = w^{(i)}_{n+1} / \sum_{j=1}^{M} w^{(j)}_{n+1}$$
  - $ **Correction** :
  - Poser
  $$\hat{p}(x^a_{n+1}|y_{0:n+1}) = \sum_{i=1}^{M} \tilde{w}^{(i)}_{n+1} K^X_{h_M}(x^a_{n+1} - \tilde{x}^a_{n+1^-}{}^{(i)})$$
  - Pour $i = 1, \ldots, M$, faire :
    - Générer $\tilde{x}^a_{n+1}{}^{(i)} \sim \hat{p}(x^a_{n+1}|y_{0:n+1})$

Par convention, nous posons $p(x^a_0|\tilde{x}^a_{-1}{}^{(i)}) = p(x^a_0)$ pour tout $i \in \{1, \ldots, M\}$. Le fait de commencer l'itération à $n = -1$ permet de prendre en compte la première observation Y_0 dans l'algorithme. L'étape $n = -1$ sera appelée étape d'initialisation par la suite.

Si le noyau $K^X_{h^X_M}$ est la densité d'une loi normale centrée, alors on peut obtenir un estimateur MMSE de X^a_n par :

$$\hat{X}^a_n = E[X^a_n|Y_{0:n}] \approx \sum_{i=1}^{M} \tilde{w}^{(i)}_n \tilde{x}^{a\,(i)}_n. \tag{6.23}$$

6.3.4 Filtre particulaire Rao-Blackwellisé

Dans cette section, nous nous intéressons uniquement aux modèles à saut de Markov dont les équations d'état sont données par le système (6.6). Nous avons vu dans la section 6.3.2 que la méthode de filtrage particulaire pouvait être employée dans le cadre des modèles à saut de Markov en considérant le vecteur (X^a_n, C_n) comme vecteur d'état caché à l'instant n. Dans ce cas, l'espace des états (qui est aussi l'espace d'évolution des particules) est donné par $\mathcal{X}^a \times \mathcal{C}$. Plus l'espace des états est grand, plus le nombre de particules nécessaire pour permettre la convergence de l'algorithme doit l'être aussi (voir la section 7.3). Dans le cadre des modèles à saut de Markov, il est possible d'améliorer les performances de l'algorithme en découplant les vecteurs aléatoires X^a_n et C_n ce qui permet de réduire l'espace des états à $\mathcal{C}$. C'est le principe même du filtre particulaire Rao-Blackwellisé. Ce filtre est notamment utilisé en poursuite de cibles (Säkkä et al. (2004)), télécommunication (Chen et al. (2000)), détection de défaillance (Flores-Quintanilla et al. (2005)), géoscience (Baziw (2005)) et traitement de l'image (Wu et al. (2003)).

L'objectif est d'approcher $p(x^a_n|y_{0:n})$ en découplant les variables X^a_n et C_n :

$$p(x^a_n|y_{0:n}) = \sum_{c_{0:n}\in\mathcal{C}^{n+1}} p(x^a_n, c_{0:n}|y_{0:n}) = \sum_{c_{0:n}\in\mathcal{C}^{n+1}} p(x^a_n|c_{0:n}, y_{0:n}) P(c_{0:n}|y_{0:n}). \tag{6.24}$$

Le problème initial peut donc se décomposer en deux sous-problèmes :

- $P(c_{0:n}|y_{0:n})$ sera approchée par filtrage particulaire ;
- $p(x^a_n|c_{0:n}, y_{0:n})$ sera approchée par une gaussienne dont l'espérance et la matrice de covariance seront déterminées par une itération du filtre de Kalman sans parfum (cf section 6.3.1).

• **Filtrage Particulaire de $\mathbf{P(c_{0:n}|y_{0:n})}$** : l'objectif est de fournir une approximation de $P(c_{0:n}|y_{0:n})$ (nous conservons le vocabulaire introduit pour l'échantillonnage séquentiel dans la section 6.3.2 malgré que $C_{0:n}$ soit à valeurs discrètes). Tout comme dans la section 6.3.2, on choisit une densité d'importance $\pi_n(c_{0:n}|y_{0:n})$, plus simple à simuler que $P(c_{0:n}|y_{0:n})$, se décomposant de la façon suivante :

$$\pi_{n+1}(c_{0:n+1}|y_{0:n+1}) \overset{\text{def}}{=} \pi_n(c_{0:n}|y_{0:n}) q_{n+1}(c_{n+1}|c_n) = q_0(c_0) \prod_{k=0}^{n} q_{k+1}(c_{k+1}|c_k) \tag{6.25}$$

avec $q_{k+1}(c_{k+1}|c_k)$ la densité de transition d'importance permettant de propager les particules de l'étape k à l'étape $k+1$ et $q_0(c_0)$ la densité d'importance permettant de

simuler C_0. La loi de $C_{0:n}$ conditionnellement à $Y_{0:n}$ est alors approchée par la distribution suivante :

$$C_{0:n} \sim \sum_{i=1}^{M} \tilde{w}_n(\tilde{c}_{0:n}^{(i)}) \delta_{\tilde{c}_{0:n}^{(i)}} \tag{6.26}$$

avec $\tilde{c}_{0:n}^{(i)}$, pour $i = 1, \ldots, M$, M réalisations de $C_{0:n}$ tirées selon la loi associée à $\pi_n(c_{0:n}|y_{0:n})$ et $\tilde{w}_n$ la fonction poids normalisée associée à $w_n : c \mapsto p(c|y_{0:n})/\pi_n(c|y_{0:n})$. Les poids peuvent être aussi calculés séquentiellement en utilisant la formule de Bayes :

$$P(c_{0:n+1}|y_{0:n+1}) = P(c_{0:n}|y_{0:n}) \frac{p(y_{n+1}|y_{0:n}, c_{0:n+1}) p(c_{n+1}|c_n)}{p(y_{n+1}|y_{0:n})}.$$

N.B. 6.10 On peut noter que, dans l'équation précédente, $p(y_{n+1}|y_{0:n}, c_{0:n+1})$ ne se simplifie pas en $p(y_{n+1}|c_{n+1})$ comme c'était le cas pour le filtre particulaire de la section 6.3.2. Ceci est dû au fait qu'on ne conditionne plus par rapport à X_{n+1}^a mais par rapport à C_{n+1} (voir l'annexe C.3).

On obtient la formule de récurrence suivante pour le calcul des poids :

$$w_{n+1}(\tilde{c}_{0:n+1}^{(i)}) = \frac{p(\tilde{c}_{0:n+1}^{(i)}|y_{0:n+1})}{\pi_{n+1}(\tilde{c}_{0:n+1}^{(i)}|y_{0:n+1})} = w_n(\tilde{c}_{0:n}^{(i)}) \frac{p(y_{n+1}|y_{0:n}, \tilde{c}_{0:n+1}^{(i)}) p(\tilde{c}_{n+1}^{(i)}|\tilde{c}_n^{(i)})}{p(y_{n+1}|y_{0:n}) q_{n+1}(\tilde{c}_{n+1}^{(i)}|\tilde{c}_n^{(i)})}. \tag{6.27}$$

$p(y_{n+1}|y_{0:n}, \tilde{c}_{0:n+1}^{(i)})$ sera calculé dans l'étape du filtre de Kalman (voir le *Nota Bene* 6.11). $p(y_{n+1}|y_{0:n})$ est une constante de normalisation des poids qu'il n'est pas nécessaire de calculer. En effet, elle est identique à toutes les trajectoires et disparaît donc lors de la normalisation des poids :

$$\tilde{w}_{n+1}(\tilde{c}_{0:n+1}^{(i)}) = w_{n+1}(\tilde{c}_{0:n+1}^{(i)}) / \sum_{j=1}^{M} w_{n+1}(\tilde{c}_{0:n+1}^{(j)}) \qquad \text{pour} \quad i = 1, \ldots, M. \tag{6.28}$$

En combinant les équations (6.26) et (6.24), on obtient une approximation de $p(x_{0:n}^a|y_{0:n})$:

$$p(x_n^a|y_{0:n}) = \sum_{c_{0:n} \in \mathcal{C}^{n+1}} p(x_n^a|c_{0:n}, y_{0:n}) P(c_{0:n}|y_{0:n}) \approx \sum_{i=1}^{M} \tilde{w}_n(\tilde{c}_{0:n}^{(i)}) p(x_n^a|\tilde{c}_{0:n}^{(i)}, y_{0:n}). \tag{6.29}$$

• **Filtrage de Kalman sans parfum de $\mathbf{p(x_n^a|\tilde{c}_{0:n}^{(i)}, y_{0:n})}$** : pour $i = 1, \ldots, M$, on approche $p(x_n^a|\tilde{c}_{0:n}^{(i)}, y_{0:n})$ par une loi normale d'espérance $\hat{x}_{n|n}^{a\ (i)}$ et de matrice de covariance $\hat{\Sigma}_{n|n}^{x^a\ (i)}$ déterminées par :

$$\begin{aligned} \hat{x}_{n|n}^{a\ (i)} &= E[X_n^a | Y_{0:n} = y_{0:n}, C_{0:n} = \tilde{c}_{0:n}^{(i)}] \\ \hat{\Sigma}_{n|n}^{x^a\ (i)} &= E[(X_n^a - \hat{x}_{n|n}^a)^T (X_n^a - \hat{x}_{n|n}^a) | Y_{0:n} = y_{0:n}, C_{0:n} = \tilde{c}_{0:n}^{(i)}]. \end{aligned} \tag{6.30}$$

Pour chaque particule $i \in \{1, \ldots, M\}$, $\hat{x}_{n|n}^{a\ (i)}$ et $\hat{\Sigma}_{n|n}^{x^a\ (i)}$ sont estimées par une itération de filtre de Kalman sans parfum. On utilise ainsi les équations décrites dans la section

6.3.1 pour calculer les différentes quantités qui interviennent dans une itération du filtre (c'est-à-dire $\hat{x}^a_{n|n-1}{}^{(i)}$, $\hat{\Sigma}^{x^a}_{n|n-1}{}^{(i)}$, $\hat{y}_{n|n-1}{}^{(i)}$, $\hat{\Sigma}^{y}_{n|n-1}{}^{(i)}$ et $\hat{\Sigma}^{x^a,y}_{n|n-1}{}^{(i)}$) et en déduire $\hat{x}^a_{n|n}{}^{(i)}$ et $\hat{\Sigma}^{x^a}_{n|n}{}^{(i)}$.

A la fin de l'étape n, on obtient donc un estimateur de $p(x^a_n|y_{0:n})$ donné par :

$$p(x^a_n|y_{0:n}) \approx \sum_{i=1}^{M} \tilde{w}_n(\tilde{c}^{(i)}_{0:n}) \mathcal{N}(x^a_n; \hat{x}^a_{n|n}{}^{(i)}, \hat{\Sigma}^{x^a}_{n|n}{}^{(i)}). \tag{6.31}$$

On obtient alors un estimateur MMSE $\hat{x}^a_n$ de X^a_n sachant $Y_{0:n}$:

$$\hat{x}^a_n = E[X^a_n|Y_{0:n}] \approx \sum_{i=1}^{M} \tilde{w}_n(\tilde{c}^{(i)}_{0:n}) \hat{x}^a_{n|n}{}^{(i)} \tag{6.32}$$

N.B. 6.11 En ce qui concerne le calcul des poids, en suivant l'approximation gaussienne donnée par le filtre de Kalman sans parfum, on approche $p(y_{n+1}|y_{0:n}, \tilde{c}^{(i)}_{0:n+1})$ par une loi normale d'espérance $\hat{y}_{n+1|n}{}^{(i)}$ et de matrice de covariance $\hat{\Sigma}^{y}_{n+1|n}{}^{(i)}$. L'équation (6.27) devient :

$$\tilde{w}_{n+1}(\tilde{c}^{(i)}_{0:n+1}) = \tilde{w}_n(\tilde{c}^{(i)}_{0:n}) \frac{\mathcal{N}(y_{n+1}; \hat{y}_{n+1|n}{}^{(i)}, \hat{\Sigma}^{y}_{n+1|n}{}^{(i)}) p(\tilde{c}^{(i)}_{n+1}|\tilde{c}^{(i)}_n)}{q_{n+1}(\tilde{c}^{(i)}_{n+1}|\tilde{c}^{(i)}_n)}$$

N.B. 6.12 On suppose que l'état initial X^a_0 suit une loi normale d'espérance x^a_0 et de matrice de covariance $\Sigma^{x^a}_0$. x^a_0 et $\Sigma^{x^a}_0$ sont supposés connus.

Algorithme 5 : Filtre Particulaire Rao-Blackwellisé

- **Initialisation**
- Pour $i = 1, \ldots, M$, faire :
  - Générer $\tilde{c}^{(i)}_0 \sim q_0(c_0)$
  - Poser $\hat{x}^a_{0|0}{}^{(i)} = x_0$ et $\hat{\Sigma}^{x^a}_{0|0}{}^{(i)} = \Sigma^x_0$
- **Itération**
- Pour $n = 0, \ldots, N-1$, faire :
  - $ **Prolongement des trajectoires** :
  - Pour $i = 1, \ldots, M$, faire :
    - Générer $\tilde{c}^{(i)}_{n+1} \sim q_{n+1}(c_{n+1}|\tilde{c}^{(i)}_n)$
    - Calculer $\hat{y}_{n+1|n}{}^{(i)}$, $\hat{\Sigma}^{y}_{n+1|n}{}^{(i)}$, $\hat{x}^a_{n+1|n+1}{}^{(i)}$ et $\hat{\Sigma}^{x^a}_{n+1|n+1}{}^{(i)}$ avec une itération de filtre de Kalman sans parfum à partir de $\hat{x}^a_{n|n}{}^{(i)}$ et $\hat{\Sigma}^{x^a}_{n|n}{}^{(i)}$
    - Calculer les poids :

$$w_{n+1}(\tilde{c}^{(i)}_{0:n+1}) = \tilde{w}_n(\tilde{c}^{(i)}_{0:n}) \frac{\mathcal{N}(y_{n+1}; \hat{y}_{n+1|n}{}^{(i)}, \hat{\Sigma}^{y}_{n+1|n}{}^{(i)}) p(\tilde{c}^{(i)}_{n+1}|\tilde{c}^{(i)}_n)}{q_{n+1}(\tilde{c}^{(i)}_{n+1}|\tilde{c}^{(i)}_n)}$$

  - $ **Normalisation des poids** :

- Pour $i = 1, \ldots, M$, faire :
  - Poser
$$\tilde{w}_{n+1}(\tilde{c}^{(i)}_{0:n+1}) = w_{n+1}(\tilde{c}^{(i)}_{0:n+1}) / \sum_{j=1}^{M} w_{n+1}(\tilde{c}^{(j)}_{0:n+1})$$

$ **Rééchantillonnage** :
- Calculer $N_{\text{eff}} = \left[\sum_{i=1}^{M} (\tilde{w}_{n+1}(\tilde{c}^{(i)}_{0:n+1}))^2\right]^{-1}$
- Si $N_{\text{eff}} \leq \eta$:
  - Pour $i = 1, \ldots, M$, faire :
    - Tirer avec remise une nouvelle trajectoire $(\tilde{c}^{(i)}_{0:n+1}, \hat{x}^{a}_{n+1|n+1}{}^{(i)}, \hat{\Sigma}^{x^a}_{n+1|n+1}{}^{(i)})$ parmi $\{(\tilde{c}^{(i)}_{0:n+1}, \hat{x}^{a}_{n+1|n+1}{}^{(i)}, \hat{\Sigma}^{x^a}_{n+1|n+1}{}^{(i)}), i \in \{1, \ldots, M\}\}$ selon une loi multinomiale $\mathscr{M}(1, \tilde{w}_n(\tilde{c}^{(1)}_{0:n+1}), \ldots, \tilde{w}_n(\tilde{c}^{(M)}_{0:n+1}))$
    - Poser $\tilde{w}^{(i)}_{n+1} = 1/M$
- Sinon, conserver le jeu de trajectoires avec leurs poids respectifs

---

Par soucis de clarté, l'observation Y_0 n'est pas prise en compte dans l'algorithme. Cependant, il est possible de l'intégrer en combinant les remarques faites pour le filtre de Kalman sans parfum et le filtre particulaire à ce sujet.

Chapitre 7

Estimation des paramètres liés au fonctionnement de la plante

L'objectif de ce chapitre est l'estimation du vecteur de paramètres Θ_{fonc} associé au fonctionnement de la plante pour le modèle de croissance $\mathcal{M}$ (voir le chapitre 2). A titre de rappel, le fonctionnement est l'ensemble des processus de création de biomasse par photosynthèse et de son allocation (voir les sections 2.2.1 et 2.2.2).

Afin d'estimer Θ_{fonc}, nous souhaitons utiliser les méthodes d'inférence bayésienne du chapitre 6. Pour cela, il faut auparavant écrire le métamodèle $\mathcal{M}$ sous la forme d'un modèle de Markov caché ou d'un modèle à saut de Markov. L'écriture d'un modèle de croissance de plante sous la forme d'un système dynamique n'est pas une idée nouvelle. Cournède et al. (2011) écrivent par exemple le système dynamique (sans bruits de modélisation) associé à une famille de modèles de plante structure-fonction caractérisée par un modèle d'allocation de carbone particulier. Afin d'estimer les paramètres du modèle, ils utilisent la méthode des moindres carrés généralisés (voir *2-stage Aitken Estimator*, Taylor (1977)). Cette approche trouve cependant ses limites lorsqu'il s'agit de caractériser les incertitudes liées aux paramètres. Trevezas and Cournède (2011) ont écrit un modèle de Markov caché pour le modèle GreenLab 1 en introduisant un cadre de bruits de modélisation et de mesure plus général. Cependant, les temps d'expansion et d'activité sont supposés tous égaux. Les paramètres ainsi que les bruits de modélisation et de mesure sont estimés par la méthode du maximum de vraissemblance via une variante stochastique de l'algorithme *Expectation-Maximization*. Dans cette section, nous proposons une troisième approche reposant sur les techniques d'inférence bayésienne du chapitre 6, le but étant l'estimation de Θ_{fonc} et de sa distribution *a posteriori*. Contrairement à Trevezas and Cournède (2011), les temps d'activité et d'expansion sont quelconques. L'approche bayésienne a déjà été employée par Wallach et al. (2006) dans le cadre des plantes mais pour d'autres familles de modèles. L'un des objectifs de ce chapitre est aussi de pouvoir comparer et critiquer les différentes méthodes sur un ensemble de cas-tests. Gaucherel et al. (2008) ont déjà effectué ce genre de comparaison concernant d'autres méthodes (moindres carré, MCMC et filtrage particulaire) pour un modèle de croissance de plante de type *process-based*.

Dans la section 7.1, nous écrivons le modèle statistique associé à $\mathcal{M}$. Nous y décrivons en particulier la nature des données disponibles ainsi que les modèles de Markov

cachés et modèles à saut de Markov associés au système dynamique. Il est donc possible d'employer les méthodes d'inférence bayésienne du chapitre 6 pour l'estimation de Θ_{fonc}. La section 7.2 donne tous les éléments permettant d'implémenter les méthodes en pratique. Ensuite, les différentes méthodes sont analysées et comparées entre elles à partir d'un ensemble de cas-tests basés sur le modèle GreenLab, cf section 7.3. La section 7.4 propose des estimateurs pour les bruits de modélisation et de mesure ainsi qu'une procédure permettant d'estimer la distribution *a posteriori* de Θ_{fonc} par bootstrap paramétrique. La dernière section est une application sur données réelles : la betterave sucrière.

N.B. 7.1 Nous supposons que les paramètres du modèle de développement $\mathcal{S}$ sont connus ainsi que les temps d'activité T_{act} des feuilles et les temps d'expansion T^o_{exp} des organes de type $o \in O$.

N.B. 7.2 Les conventions de notation du chapitre 6 sont encore valables dans ce chapitre.

7.1 Description du modèle statistique

Dans cette section, nous posons un cadre statistique s'inspirant des travaux de Trevezas and Cournède (2011) et qui nous permettra d'utiliser les méthodes d'inférence bayésienne décrites dans le chapitre 6. Nous nous intéressons d'abord aux données botaniques que l'on peut récolter à partir d'une plante et qui nous sont nécessaires pour l'estimation du vecteur de paramètres Θ_{fonc} (voir section 7.1.1). Ensuite, nous écrivons le modèle de croissance de plante sous la forme d'un modèle de Markov caché ou d'un modèle à saut de Markov suivant si l'évolution de la structure de la plante donnée par $\{N_n\}_{n\geq 0}$ (voir la section 2.1.6) est déterministe ou stochastique, cf section 7.1.2.

7.1.1 Les données botaniques

Supposons que l'on dispose d'une plante et que l'on est capable de donner l'âge (en cycles de développement) de chacun des organes la composant. A un instant $T_{obs} \in \mathbb{N}^*$, la plante est découpée et un certain nombre d'organes sont pesés. Pour tout $(o, n) \in \mathcal{O} \times \{0, \dots, T_{obs}\}$, notons $\mathscr{N}^o_n$ le nombre d'organes de type o créés au cycle de développement n qui ont été pesés. Etant donné que le nombre total d'organe de cette sorte vaut $N^o_{n+1} - N^o_n$ (cf section section 2.1.6), nous avons alors immédiatement $0 \leq \mathscr{N}^o_n \leq N^o_{n+1} - N^o_n$. Rappelons que la masse $M^o_{n,k}$ d'un organe de type o d'âge k au cycle de développement n est donnée par la relation (voir l'équation (2.4) du chapitre 2) :

$$M^o_{n,k} = \sum_{l=0}^{\min(k, T^o_{exp})-1} Al^{o,l}_{n-k+l}(Q_{n-k+l}, N_{0\to n-k+l+1}, \Theta_{fonc}), \qquad 1 \leq n \text{ et } 1 \leq k \leq n. \quad (7.1)$$

Au moment de l'observation expérimentale (c'est-à-dire à la fin du cycle de développement T_{obs}), un organe de type o créé au cycle n a un âge $k = T_{obs} + 1 - n$. Sa masse est donc $M^o_{T_{obs}+1, T_{obs}+1-n}$ (on considère que l'âge d'un organe à la toute fin du cycle T_{obs}

est le même qu'au tout début du cycle $T_{obs}+1$). Si $\mathscr{N}_n^o \geq 1$, notons $M^o_{T_{obs}+1,T_{obs}+1-n}(i)$ la masse du $i$ème organe pesé avec $i = 1, \ldots, \mathscr{N}_k^o$. Chacune de ces masses est entâchée d'une erreur de mesure additive $\epsilon_n^o(i)$ qui est due au protocole de recueil des données (imprécisions liées à l'expérimentateur, aux appareils de mesure ...). Nous supposons que toutes les erreurs de mesure sont mutuellement indépendantes deux à deux, ne dépendent que du type d'organe considéré et suivent une loi normale centrée de variance σ_o^2 (cette variance est supposée indépendante du cycle de création n). La masse $M^o_{T_{obs}+1,T_{obs}+1-n}(i)$ est donc donnée par :

$$M^o_{T_{obs}+1,T_{obs}+1-n}(i) = \sum_{l=0}^{\min(T_{obs}+1-n,T^o_{exp})-1} Al^{o,l}_{n+l}(Q_{n+l}, N_{0\to n+1+l}, \Theta_{fonc}) + \epsilon_n^o(i), \;\; i = 1, \ldots, \mathscr{N}_n^o. \tag{7.2}$$

Soit $\overline{M^o_{T_{obs}+1,T_{obs}+1-n}}$ la valeur moyenne des $\left\{M^o_{T_{obs}+1,T_{obs}+1-n}(i)\right\}_i$ pour un type d'organe o donné :

$$\overline{M^o_{T_{obs}+1,T_{obs}+1-n}} = \begin{cases} \frac{1}{\mathscr{N}_n^o}\sum_{i=1}^{\mathscr{N}_n^o} M^o_{T_{obs}+1,T_{obs}+1-n}(i) & \text{si } \mathscr{N}_n^o \geq 1, \\ 0 & \text{sinon.} \end{cases}$$

Par la suite, nous supposons $\mathscr{N}_n^o \geq 1$ pour tout $o \in \mathcal{O}$ et $n \in \{0, \ldots, T_{obs}\}$ afin d'alléger les notations (l'adaptation au cas $\mathscr{N}_n^o = 0$ est immédiate). En conséquence, $\overline{M^o_{T_{obs}+1,T_{obs}+1-n}}$ suit une loi normale d'espérance la moyenne des masses non bruitées et de variance $\sigma_o^2/\mathscr{N}_n^o$. On définit alors le vecteur d'observation Y_n comme le vecteur des masses moyennes des organes de $\mathcal{O}$ qui sont apparus au cycle de développement n :

$$Y_n = \left(\overline{M^o_{T_{obs}+1,T_{obs}+1-n}}\right)_{o\in\mathcal{O}}. \tag{7.3}$$

En développant Y_n grâce à l'équation (7.2), nous avons :

$$Y_n = \left(\frac{1}{\mathscr{N}_n^o} \sum_{i=1}^{\mathscr{N}_n^o} \sum_{l=0}^{\min(T_{obs}+1-n,T^o_{exp})-1} Al^{o,l}_{n+l}(Q_{n+l}, N_{0\to n+1+l}, \Theta_{fonc})(i) \right)_{o\in\mathcal{O}} + V_n \tag{7.4}$$

avec $V_n = ((1/\mathscr{N}_n^o)\sum_{i=1}^{\mathscr{N}_n^o} \epsilon_n^o(i))_{o\in\mathcal{O}}$ un vecteur gaussien centré de matrice de covariance R_n. R_n est une matrice diagonale dont les éléments diagonaux sont donnés par les composantes du vecteur $(\sigma_o^2/\mathscr{N}_n^o)_{o\in\mathcal{O}}$. $\{V_n\}_{n\geq 0}$ est une suite de vecteurs aléatoires i.i.d. En notant $T^m_{exp} = \max\{T^o_{exp}\}_{o\in\mathcal{O}}$ et $\mathscr{N}_n = (\mathscr{N}_n^o)_{o\in\mathcal{O}}$, Y_n peut se mettre sous la forme :

$$Y_n = m_n(Q_n, \ldots, Q_{n+T^m_{exp}-1}, N_{0\to n+T^m_{exp}}, \mathscr{N}_n, \Theta_{fonc}) + V_n. \tag{7.5}$$

avec m_n une fonction borélienne et la convention que $Q_k = 0$ et $N_{0\to k} = N_{0\to T_{obs}+1}$ si $k \geq T_{obs}+1$ (la plante est découpée à la fin du cycle de développement T_{obs}).

N.B. 7.3 Lorsque l'on travaille avec un système dynamique évoluant à temps discret, il est naturel de prendre l'observation Y_n comme étant un ensemble de mesures réalisées sur le système à l'instant n. Les mesures $\{Y_n\}_{n\geq 0}$ reflètent l'évolution du système dynamique. On peut cependant noter que ce n'est pas le cas avec le protocole de recueil de données

détaillé dans cette section. En effet, l'observation d'une plante est destructive puisque l'on pèse ses organes (il existe d'autres méthodes d'observation non destructive qui ne considèrent pas les masses mais elles ne sont pas prises en compte dans cette thèse). Le système « plante » ne peut être observé qu'une seule fois puisque les mesures sont destructives. Les données Y_n ainsi définies dans cette section ne sont pas liées à une évolution du système mais plutôt à l'architecture de la plante et plus particulièrement à l'ordre d'apparition des organes (voir Trevezas and Cournède (2011)) ce qui est une façon originale de percevoir le système dynamique.

7.1.2 Modèle de Markov caché et modèle à saut de Markov associés

Nous souhaitons poser un cadre statistique pour le modèle de croissance de plante $\mathcal{M}$. Plus particulièrement, nous allons décrire le système dynamique associé sous la forme d'un modèle de Markov caché ou d'un modèle à saut de Markov ce qui nous permettra d'utiliser les méthodes d'inférence bayésienne du chapitre 6 pour l'estimation du vecteur de paramètres Θ_{fonc}.

Le fonctionnement de la plante a été introduit comme l'ensemble des processus de création et d'allocation de biomasse. L'allocation est caractérisée par l'équation (7.5) donnant le vecteur des masses moyennes des organes créés à un cycle de développement donné. La création de biomasse est gouvernée par l'équation de photosynthèse (2.7) du chapitre 2. Cette équation permet de calculer la quantité de biomasse Q_n disponible au cycle de développement n :

$$Q_n = \Phi_n(Q_{(n-T_{act})^+}, \dots, Q_{n-1}, N_{0\to n}, E_n, \Theta_{fonc}), \qquad n \geq 1. \tag{7.6}$$

avec $(n - T_{act})^+ = \max(n - T_{act}, 0)$. Du point de vue modélisation statistique, il n'est pas raisonnable de supposer que l'équation de production soit déterministe. En effet, plusieurs sources d'incertitudes affectent le calcul de Q_n :

- incertitude de modélisation : l'équation de Beer-Lambert suppose une distribution poissonienne des feuilles dans le couvert végétal (voir Vose et al. (1995)). Le calcul de Q_n est donc le résultat d'une extrapolation.
- incertitude sur la variable environnement $E(n)$. La valeur attribuée à $E(n)$ est le résultat d'un ensemble d'observations expérimentales, chacune d'entre elles étant entâchée de sa propre erreur de mesure.

Partant de ces différents constats, il convient d'ajouter un bruit de modélisation à l'équation (7.6) noté ω_n que l'on suppose multiplicatif :

$$Q_n = \Phi_n(Q_{(n-T_{act})^+}, \dots, Q_{n-1}, N_{0\to n}, E_n, \Theta_{fonc})(1 + \omega_n), \qquad n \geq 1. \tag{7.7}$$

Nous supposons également que ω_n suit une loi gaussienne centrée de variance K_n et que les suites $\{\omega_n\}_{n\geq 0}$ et $\{V_n\}_{n\geq 0}$ (bruits de mesure) sont mutuellement indépendantes.

Les variables Q_k, $k \geq 0$, jouent donc un rôle clé dans la croissance de la plante et le processus $(Q_k)_{k\geq 0}$ caractérise l'évolution du système dynamique. De plus, ces variables

ne sont pas directement observables. Elles le sont par l'intermédiaire des masses des différents organes et donc des vecteurs d'observation $\{Y_n\}_{n\geq 0}$. Il apparaît donc naturel de définir l'état caché du système dynamique à partir de l'ensemble $\{Q_k\}_{n\geq 0}$. Pour cela, nous introduisons $T \in \mathbb{N}^*$, le maximum des temps d'activité et d'expansion de la plante :

$$T = \max\left(T_{act}, T^m_{exp}\right).$$

Nous définissons alors le vecteur d'état caché du système dynamique de la façon suivante :

$$X_n = (Q_{n+T^m_{exp}-T}, \ldots, Q_{n+T^m_{exp}-1}), \qquad n \geq 0,$$

avec la convention $Q_k = 0$ si $k < 0$.

N.B. 7.4 Les indices de la variable Q dans X_n ont été choisis de sorte que le système « plante » puisse se mettre sous la forme d'un modèle de Markov caché ou d'un modèle à saut de Markov.

L'observation du système à l'instant n est donnée par le vecteur Y_n des masses moyennes des organes créés au cycle de développement n (voir l'équation (7.3)). L'équation de production (7.7) permet de créer X_{n+1} à partir de X_n et sera donc liée à l'équation d'évolution du système dynamique. L'équation (7.5) donne la masse des organes en fonction de X_n et sera alors associée au modèle de mesure. Il est donc possible d'écrire complètement le modèle statistique associé à $\mathcal{M}$. Deux situations sont possibles : l'évolution de la structure de la plante est déterministe (*i.e.* nous connaissons N_n pour tout $n \geq 0$) ou stochastique ($(N_n)_{n\geq 0}$ est un processus de branchement de Galton-Watson multitype, cf le théorème 3.3.3 du chapitre 3).

Théorème 7.1.1 *Supposons que le processus $(N_n)_{n\geq 0}$ soit déterministe et l'état initial Q_0 soit connu. Dans ce cas, le modèle de croissance de plante $\mathcal{M}$ caractérisé par le processus à temps discret $\{(X_n, Y_n)\}_{n\geq 0}$ censuré à l'instant T_{obs} est un modèle de Markov caché dont les modèles d'évolution et d'observation sont donnés par les densités de probabilité suivantes :*

$$\begin{cases} p(x'_{n+1}|x_n) = \mathcal{N}\left(q'_{n+T^m_{exp}}; \phi_{n+T^m_{exp}}, K_{n+T^m_{exp}}\left(\phi_{n+T^m_{exp}}\right)^2\right) \displaystyle\prod_{i=n+T^m_{exp}-T+1}^{n+T^m_{exp}-1} \delta_{q_i}(q'_i) \\ p(y_n|x_n) = \mathcal{N}\left(y_n; m_n(q_n, \ldots, q_{n+T^m_{exp}-1}, N_{0\to n+T^m_{exp}}, \mathscr{N}_n, \Theta_{fonc}), R_n\right). \end{cases}$$

avec :

$$\phi_{n+T^m_{exp}} = \Phi_{n+T^m_{exp}}(q_{(n+T^m_{exp}-T_{act})^+}, \ldots, q_{n+T^m_{exp}-1}, N_{0\to n+T^m_{exp}}, E_{n+T^m_{exp}}, \Theta_{fonc}),$$

$x_n = (q_{n+T^m_{exp}-T}, \ldots, q_{n+T^m_{exp}-1})$ *et δ_u le symbole de Kronecker centré en $u \in \mathbb{R}$.*

Preuve Nous allons écrire le système dynamique sous la forme du modèle de Markov caché donné par la définition 6.2.1 du chapitre 6. Notons $d_n = T + d_{\omega_{n+T^m_{exp}-1}} + d_{\Theta_{fonc}}$ et introduisons tout d'abord la fonction f_n de $\mathbb{R}^{d_n}$ dans $\mathbb{R}^T$ définie de la façon suivante :

$$\forall X_{n-1} = (Q_{n+T^m_{exp}-T-1}, \ldots, Q_{n+T^m_{exp}-2}) \in \mathbb{R}^T,\ \omega_{n+T^m_{exp}-1} \in \mathbb{R}^{d_{\omega_{n+T^m_{exp}-1}}},\ \Theta_{fonc} \in \mathbb{R}^{d_{\Theta_{fonc}}}, \tag{7.8}$$

$$f_n(X_{n-1}, \omega_{n+T^m_{exp}-1}, \Theta_{fonc}) = (Q_{n+T^m_{exp}-T}, \ldots, Q_{n+T^m_{exp}-2}, A_n(X_{n-1}, \omega_{n+T^m_{exp}-1}, \Theta_{fonc}))$$

avec :

$$A_n(X_{n-1}, \omega_{n+T^m_{exp}-1}, \Theta_{fonc}) = \\ \Phi_{n+T^m_{exp}-1}(Q_{(n+T^m_{exp}-T_{act}-1)^+}, \ldots, Q_{n+T^m_{exp}-2}, N_{0\to n+T^m_{exp}-1}, E_{n+T^m_{exp}-1}, \Theta_{fonc})(1+\omega_{n+T^m_{exp}-1}) \tag{7.9}$$

De même, introduisons la fonction g_n de $\mathbb{R}^{T+d_{\Theta_{fonc}}}$ dans $\mathbb{R}^{dim(\mathcal{O})}$ définie de la façon suivante :

$$\forall X_n = (Q_{n+T^m_{exp}-T}, \ldots, Q_{n+T^m_{exp}-1}) \in \mathbb{R}^T,\ \Theta_{fonc} \in \mathbb{R}^{d_{\Theta_{fonc}}}, \\ g_n(X_n, \Theta_{fonc}) = m_n(Q_n, \ldots, Q_{n+T^m_{exp}-1}, N_{0\to n+T^m_{exp}}, \mathcal{N}_n, \Theta_{fonc}). \tag{7.10}$$

Dans ce cas, le processus discret $\{(X_n, Y_n)\}_{n\geq 0}$ vérifie le système d'équations suivant :

$$\begin{cases} X_{n+1} = f_{n+1}(X_n, \omega_{n+T^m_{exp}}, \Theta_{fonc}), \\ Y_n = g_n(X_n, \Theta_{fonc}) + V_n. \end{cases}$$

En posant $W_n = \omega_{n+T^m_{exp}-1}$ et en remarquant que $\{W_n\}_{n\geq 0}$ et $\{V_n\}_{n\geq 0}$ des suites mutuellement indépendantes de vecteurs aléatoires i.i.d., nous avons bien que $\{(X_n, Y_n)\}_{n\geq 0}$ est un modèle de Markov caché d'après la définition 6.2.1. Les densités de probabilité associées au modèle d'évolution et d'observation découlent directement du fait que W_n et V_n suivent des lois normales centrées de matrice de covariance respective $K_{n+T^m_{exp}-1}$ et R_n.

□

N.B. 7.5 La densité associée à la loi de X_0 ne peut pas s'exprimer de façon explicite. Cependant, elle est facile à simuler (voir la section 7.2.1).

Dans le cas où l'évolution de la structure de la plante est stochastique, le processus $\{(X_n, Y_n)\}_{n\geq 0}$ n'est plus un modèle de Markov caché mais un modèle à saut de Markov. Considérons le processus $\{C_n\}_{n\geq 0}$ défini de la façon suivante :

$$C_n = \left(N_{0\to n+T^m_{exp}}, \mathbf{0}_{(T_{obs+1-n-T^m_{exp}})card(N_0)}\right), \qquad n \geq 0.$$

avec $\mathbf{0}_k$ le vecteur de taille k dont toutes les composantes sont nulles. Etant donné que $\{N_n\}_{n\geq 0}$ est un processus de branchement de Galton-Watson multitype (voir le théorème 3.3.3 du chapitre 3), $\{C_n\}_{n\geq 0}$ est donc une chaîne de Markov caractérisée par sa distribution initiale $P(c_0)$ et ses probabilités de transition $P(c_{n+1}|c_n) = P(N_{n+T^m_{exp}+1}|N_{n+T^m_{exp}})$.

N.B. 7.6 Nous conservons les majuscules dans $P(N_{n+T^m_{exp}+1}|N_{n+T^m_{exp}})$ pour éviter toute confusion entre le vecteur aléatoire N et l'indice n utilisé à maintes reprises dans la thèse.

La distribution initiale $P(c_0)$ n'est pas exprimable explicitement mais est facilement simulable (voir la section 7.2.1). Nous avons alors (voir Loi et al. (2011)) :

Théorème 7.1.2 *Supposons que $(N_n)_{n\geq 0}$ soit un processus aléatoire et que l'état initial Q_0 soit connu. Dans ce cas, le modèle de croissance de plante $\mathcal{M}$ caractérisé par le processus à temps discret $\{(X_n, Y_n)\}_{n\geq 0}$ censuré à l'instant T_{obs} est un modèle à saut de Markov dont les modèles d'évolution et d'observation sont donnés par les densités de probabilité suivantes :*

$$\begin{cases} p(x'_{n+1}|x_n, c_{n+1}) = \mathcal{N}\left(q'_{n+T^m_{exp}}; \phi_{n+T^m_{exp}}, K_{n+T^m_{exp}}\left(\phi_{n+T^m_{exp}}\right)^2\right) \prod_{i=n+T^m_{exp}-T+1}^{n+T^m_{exp}-1} \delta_{q_i}(q'_i) \\ p(y_n|x_n, c_n) = \mathcal{N}\left(y_n; m_n(q_n, \ldots, q_{n+T^m_{exp}-1}, c_n, \mathcal{N}_n, \Theta_{fonc}), R_n\right). \end{cases}$$

avec :

$$\phi_{n+T^m_{exp}} = \Phi_{n+T^m_{exp}}(q_{(n+T^m_{exp}-T_{act})^+}, \ldots, q_{n+T^m_{exp}-1}, c_{n+1}, E_{n+T^m_{exp}}, \Theta_{fonc}),$$

$x_n = (q_{n+T^m_{exp}-T}, \ldots, q_{n+T^m_{exp}-1})$ *et δ_u le symbole de Kronecker centré en $u \in \mathbb{R}$.*

Preuve La démonstration est immédiate. Introduisons les mêmes fonctions f_n et g_n définies respectivement par les équations (7.8) et (7.10) en prenant soin cette fois de marquer la dépendance en $N_{0\to k}$ avec $k \geq 0$. Nous avons alors immédiatement que le processus discret $\{(X_n, Y_n)\}_{n\geq 0}$ vérifie le système d'équations suivant :

$$\begin{cases} X_{n+1} = f_{n+1}(X_n, C_{n+1}, W_{n+1}, \Theta_{fonc}), \\ Y_n = g_n(X_n, C_n, \Theta_{fonc}) + V_n. \end{cases}$$

avec W_n pour $n \geq 0$ défini de la même façon que dans la démonstration du théorème 7.1.1. D'après la définition 6.2.2 du chapitre 6, nous avons bien que $\{(X_n, Y_n)\}_{n\geq 0}$ est un modèle à saut de Markov. Les densités de probabilité associées au modèle d'évolution et d'observation découlent également du fait que W_n et V_n suivent des lois normales centrées de matrice de covariance respective $K_{n+T^m_{exp}-1}$ et R_n.

N.B. 7.7 X_{n+1} dépend en fait de C_n. Etant donné que toute l'information de C_n est contenue dans C_{n+1}, X_{n+1} dépend également de C_{n+1}. C'est cette dépendance que nous choisissons afin de représenter le système sous la forme d'un modèle à saut de Markov.

□

7.2 Mise en œuvre pratique des méthodes d'inférence bayésienne

Suivant que l'évolution de la structure de la plante est déterministe ou non, le modèle de croissance de plante $\mathcal{M}$ peut être décrit soit par un modèle de Markov caché soit par un modèle à saut de Markov. Il est donc possible d'utiliser les méthodes d'inférence bayésienne du chapitre 6 pour estimer le vecteur de paramètres Θ_{fonc}. Dans cette section, nous expliquons comment mettre en pratique ces méthodes. Nous montrons comment adapter les algorithmes et donnons quelques conseils quant à leur utilisation.

Les différents points mentionnés par la suite sont une combinaison de résultats empruntés à la littérature scientifique sur le sujet (voir Cappé et al. (2005), Caron (2006)) et de résultats observés par le biais de l'expérimentation durant la thèse. Pour la suite de cette section, nous supposons connues ou fixées la biomasse Q_0 contenue dans la graine ainsi que les variances K_n et matrices de covariance R_n, $n \in \mathbb{N}$, associées respectivement aux bruits de modélisation et aux bruits de mesure (voir la section 7.3.4 concernant ce sujet).

7.2.1 Initialisation des algorithmes

• **Initialisation pour le filtre de Kalman sans parfum** : nous donnons une méthode permettant de construire les conditions initiales x_0^a et $\Sigma_0^{x^a}$. Pour cela, il faut en premier lieu attribuer une valeur de départ Θ_{fonc}^{init} au vecteur de paramètres Θ_{fonc}. Nous attribuons alors à chaque composante de Θ_{fonc}^{init} une valeur d'un ordre de grandeur correspondant au paramètre que cette composante représente. Par exemple, si l'on sait que le paramètre en question varie typiquement entre 1 et 10, on peut lui attribuer la valeur initiale 5. Tout autre choix dans l'intervalle $[1; 10]$ est correct également. Par contre, il ne faut pas lui attribuer une valeur de 100. Cela présuppose que nous sommes capable de donner un ordre de grandeur raisonnable pour chacun des paramètres ce qui est le cas dans la pratique pour les modèles de croissance de plante (cf section 7.3.4 pour des exemples). Pour construire x_0^a, il faut également attribuer une valeur de départ x_{init}^0 au vecteur d'état caché X_0. Afin de simplifier les calculs, nous construisons x_0^a dans le cas où $T_{exp}^m \geq T_{act}$, c'est-à-dire $T = T_{exp}^m$ (la démarche est identique pour le cas $T_{exp}^m < T_{act}$ en adaptant les indices). Dans ce cas, $X_0 = (Q_0, \ldots, Q_{T_{exp}^m - 1}) = (Q_0, \ldots, Q_{T-1})$. Construire x_0^a revient à attribuer une valeur initiale à Q_k avec $k \in \{1, \ldots, T_{exp}^m - 1\}$. Ainsi, nous aurons $x_0^a = (x_{init}^0, \Theta_{fonc}^{init})$. Rappelons qu'à la base nous ne disposons seulement que de Q_0. Soit alors le vecteur $x_{init} = (\mathbf{0}_{T-1}, Q_0)$ et $f_{1:T-1}$ la composée en X des $T-1$ fonctions $f_1, \ldots, f_{T-1}$ (la fonction f_n est définie par l'équation (7.8)) :

$$\begin{aligned} &\forall (X, \omega_{1:T-1}, \Theta), \\ &f_{1:T-1}(X, \omega_{1:T-1}, \Theta) = f_{T-1}(f_{T-2}(\ldots(f_1(X, \omega_1, \Theta), \omega_2, \Theta), \ldots, \omega_{T-2}, \Theta), \omega_{T-1}, \Theta) \end{aligned} \tag{7.11}$$

x_{init}^0 est alors choisi comme la valeur moyenne du premier état caché X_0 conditionnellement à Q_0 :

$$x_{init}^0 = \mathbb{E}[f_{1:T-1}(x_{init}, \omega_{1:T-1}, \Theta_{fonc}^{init})].$$

N.B. 7.8 Dans l'équation précédente, les seules variables aléatoires sont données par le vecteur $\omega_{1:T-1}$.

Pour calculer x_{init}^0, nous pouvons utiliser une procédure de transformation sans parfum (voir l'annexe C.2). Associons au vecteur $\omega_{1:T-1}$ son vecteur des valeurs moyennes $\mathbf{0}_{T-1}$ et sa matrice de covariance diagonale D dont les éléments diagonaux sont donnés par le vecteur $(K_1, \ldots, K_{T-1})$. Considérons les $2T-1$ sigma-points χ_ω^i associés au couple $(\mathbf{0}_{T-1}, D)$ avec leur poids respectif $\mathcal{W}^i$ (voir l'annexe C.2 pour la définition des sigma-points). Propageons ces sigma-points à travers la fonction $f_{1:T-1}$ en attribuant

respectivement aux entrées X et Θ les vecteurs x_{init} et Θ_{fonc}^{init} et nous obtenons $2T-1$ nouveaux sigma-points $\chi_{x_0}^i$ associés à l'état caché X_0, chacun d'entre eux étant de taille T :

$$\forall i \in \{1, \ldots, 2T-1\}, \qquad \chi_{x_0}^i = f_{1:T-1}(x_{init}, \chi_\omega^i, \Theta_{fonc}^{init}).$$

Le vecteur x_{init}^0 est alors donné par :

$$x_{init}^0 = \sum_{i=1}^{2T-1} \mathcal{W}^i \, \chi_{x_0}^i.$$

L'initialisation de la matrice de covariance $\Sigma_0^{x^a}$ se fait en utilisant la transformation sans parfum précédente. Nous choisissons pour $\Sigma_0^{x^a}$ la matrice diagonale par bloc dont les blocs sont donnés successivement par Σ_{init} et B avec :

$$\Sigma_{init} = \sum_{i=1}^{2T-1} \mathcal{W}^i (\chi_{x_0}^i - x_{init}^0)^t (\chi_{x_0}^i - x_{init}^0)$$

avec $(.)^t$ la transposée d'une matrice et B une matrice diagonale dont les éléments diagonaux sont égaux à 10^{-4} (le choix d'une telle matrice est également expliqué dans la section 7.3.4).

• **Simulation suivant la distribution initiale pour le filtrage particulaire et le filtrage particulaire convolé** : nous expliquons comment obtenir un ensemble de M réalisations $\{\tilde{x}_0^{a(i)}, i = 1, \ldots, M\}$ pour ces deux méthodes particulaires. Chaque réalisation $\tilde{x}_0^{a(i)}$ se décompose en un vecteur $\tilde{x}_0{}^{(i)}$ associé à l'état caché X_0 du système dynamique et un vecteur $\tilde{\Theta}_0{}^{(i)}$ associé au vecteur de paramètres Θ_{fonc}. Expliquons tout d'abord comment avoir les vecteurs $\{\tilde{\Theta}_0{}^{(i)}, i = 1, \ldots, M\}$. Ces derniers sont choisis de façon à être uniformément répartis dans l'espace des paramètres $\mathcal{P}$ (on suppose que $\mathcal{P}$ est borné, ce qui est le cas dans la pratique). Pour ce faire, on crée d'abord un ensemble de points uniformément répartis dans $\mathcal{P}$. On note M_1 son cardinal. On affecte ensuite un nombre M_2 de particules à chacun des points précédents. On obtient alors un nombre $M = M_1 M_2$ de particules. En utilisant cette procédure, on augmente les chances d'orienter directement les particules dans la bonne zone de l'espace $\mathcal{P}$. Plus le nombre de points M_1 est important, plus l'algorithme converge rapidement. Cependant, si l'espace $\mathcal{P}$ est de grande dimension, M_1 peut devenir assez conséquent. En effet, si K valeurs sont choisies pour chaque dimension de $\mathcal{P}$, on a donc $M_1 = K^{dim(\mathcal{P})}$. Par exemple, pour 10 paramètres et $K = 5$ (ce qui est une répartition raisonnable), $M1$ vaut un peu moins de dix millions. Ce phénomène est connu sous le nom d'explosion combinatoire. Si le nombre de paramètres à estimer est grand, la valeur de K doit être faible. On ne peut donc proposer qu'une répartition grossière dans ce cas. Supposons maintenant que nous disposons d'un ensemble de vecteurs de paramètres $\{\tilde{\Theta}_0{}^{(i)}, i = 1, \ldots, M\}$. Pour construire les états cachés initiaux correspondants $\{\tilde{x}_0{}^{(i)}, i = 1, \ldots, M\}$, nous partons d'un vecteur initial $x_{init} = (\mathbf{0}_{T-1}, Q_0)$. Tout comme pour le filtre de Kalman sans parfum, nous ne considérons que le cas où $T_{exp}^m \geq T_{act}$, c'est-à-dire $T = T_{exp}^m$, afin de simplifier les notations (la démarche est identique pour le cas $T_{exp}^m < T_{act}$ en adaptant les indices). Soit alors $\{\omega_{1:T-1}{}^{(i)}, i = 1, \ldots, M\}$ un ensemble de M réalisations indépendantes d'une

loi normale centrée de dimension $T-1$ dont la matrice de covariance est diagonale et les éléments diagonaux sont donnés par le vecteur $(K_1, \ldots, K_{T-1})$. Les vecteurs ${\tilde{x}_0}^{(i)}$ sont alors donnés par :

$$ {\tilde{x}_0}^{(i)} = f_{1:T-1}(x_{init}, {\omega_{1:T-1}}^{(i)}, {\tilde{\Theta}_0}^{(i)}) $$

avec $f_{1:T-1}$ la fonction donnée par l'équation (7.11). ${\tilde{x}_0}^{(i)}$ est en fait une réalisation du système dynamique caractérisé par le vecteur de paramètres ${\tilde{\Theta}_0}^{(i)}$ sur $T-1$ étapes. Nous posons alors :

$$ \forall i \in \{1, \ldots, M\}, \qquad \tilde{x}_0^{a(i)} = ({\tilde{x}_0}^{(i)}, {\tilde{\Theta}_0}^{(i)}). $$

• **Initialisation pour le filtre particulaire Rao-Blackwellisé** : l'objectif est d'obtenir un ensemble de réalisations $\{(\tilde{c}_0^{(i)}, x_0^{a(i)}, \Sigma_0^{x^a(i)}), i = 1, \ldots, M\}$. L'ensemble $\{\tilde{c}_0^{(i)}, i = 1, \ldots, M\}$ est simplement obtenu en simulant M trajectoires du processus de branchement $\{N_n\}_{n\geq 0}$ sur T_{exp}^m étapes en commençant avec $N_0 = (\ldots, \delta_{b_1}(b), \ldots)_{b\in\mathcal{B}}$. Pour chaque réalisation $\tilde{c}_0^{(i)}$, $x_0^{a(i)}$ et $\Sigma_0^{x^a(i)}$ sont déterminés de la même façon pour que l'initialisation du filtre de Kalman sans parfum.

7.2.2 Itérations multiples des algorithmes

A une étape $n \in \{0, \ldots, T_{obs}\}$ donnée, les quatre algorithmes bayésiens donnent un estimateur $\hat{p}(x_n^a|y_{0:n})$ de la densité de probabilité $p(x_n^a|y_{0:n})$. En conséquence, il est possible d'estimer l'état caché augmenté du système dynamique en utilisant l'estimateur MMSE :

$$ \hat{X}_n^a = \mathbb{E}\left[X_n^a|Y_{0:n}\right] \approx \int_{\mathcal{X}^a} x_n^a \hat{p}(x_n^a|y_{0:n}) \lambda(dx_n^a). $$

Nous obtenons donc une estimation de Θ_{fonc} en extrayant les composantes de $\hat{X}_n^a$ associées au vecteur de paramètres. Afin de bénéficier de toutes l'information fournie par les données $Y_{0:T_{obs}}$, la meilleure estimation de Θ_{fonc} est celle obtenue à partir de l'estimation du dernier vecteur d'état caché augmenté $\hat{X}_{T_{obs}}^a$. Notons alors $\hat{\Theta}_{fonc}$ cette estimation. En général, à la fin d'une itération complète des algorithmes (c'est-à-dire le déroulement des étapes 0 à T_{obs}), le vecteur $\hat{\Theta}_{fonc}$ obtenu est encore bien loin de la vraie valeur de Θ_{fonc} que l'on notera Θ_{fonc}^{vrai} par la suite. L'idée est alors d'appliquer une nouvelle fois l'algorithme en utilisant comme valeur d'entrée l'estimation $\hat{\Theta}_{fonc}$ obtenue en sortie de l'itération précédente (ou les particules associées pour les méthodes particulaires) :

• **Filtre de Kalman sans parfum** : l'initialisation de la nouvelle itération se fait de la même façon que décrit dans la section 7.2.1 sauf qu'au lieu de donner à chaque composante de Θ_{fonc}^{init} une valeur correspondant à l'ordre de grandeur de la composante en question, nous imposons $\Theta_{fonc}^{init} = \hat{\Theta}_{fonc}$. Cela permet de conserver l'information acquise au cours de l'itération précédente sur le vecteur Θ_{fonc}.

N.B. 7.9 Pour l'initialisation de la première itération, il est également possible d'utiliser un vecteur Θ_{fonc} obtenu par un estimateur classique plus facile à avoir comme, par exemple, les moindres carrés généralisés (voir Cournède et al. (2011)).

• **Filtre particulaire et le filtre particulaire convolé** : à la fin de l'étape T_{obs}, nous disposons d'un ensemble de particules $\{\tilde{x}^a_{0:T_{obs}}{}^{(i)}, i = 1, \ldots, M\}$ avec les poids normalisés associés $\{\tilde{w}_{T_{obs}}(\tilde{x}^a_{0:T_{obs}}{}^{(i)}), i = 1, \ldots, M\}$. Plus précisément, nous avons un ensemble de vecteurs de paramètres $\{\tilde{\Theta}_{T_{obs}}{}^{(i)}, i = 1, \ldots, M\}$ avec les mêmes poids normalisés en extrayant de $\tilde{x}^a_{T_{obs}}{}^{(i)}$ le vecteur $\tilde{\Theta}_{T_{obs}}{}^{(i)}$ associé à Θ_{fonc}. Ainsi, nous initialisons $\{\tilde{x}^{a(i)}_0, i = 1, \ldots, M\}$ de la même façon que dans la section 7.2.1 sauf que cette fois nous choisissons de prendre $\{\tilde{\Theta}_{T_{obs}}{}^{(i)}, i = 1, \ldots, M\}$ pour l'ensemble $\{\tilde{\Theta}_0{}^{(i)}, i = 1, \ldots, M\}$. Les poids associés aux nouvelles particules $\{\tilde{x}^{a(i)}_0, i = 1, \ldots, M\}$ sont $\{\tilde{w}_{T_{obs}}(\tilde{x}^a_{0:T_{obs}}{}^{(i)}), i = 1, \ldots, M\}$. En procédant de cette manière, nous donnons au vecteur Θ_{fonc} une distribution initiale qui tient compte des informations obtenues lors de l'itération précédente de l'algorithme ce qui va permettre une meilleure estimation du vecteur de paramètres.

• **Filtre particulaire de Rao-Blackwell** : l'initialisation de la nouvelle itération de l'algorithme se fait de la même façon que pour le filtre de Kalman sans parfum en adaptant l'estimateur MMSE pour le filtrage particulaire (cf section 6.3.4).

Pour la suite du chapitre, nous noterons $\hat{\Theta}^{(k)}_{fonc}$ l'estimation MMSE de Θ_{fonc} après la k-ième itération complète d'un des quatre algorithmes bayésiens.

7.2.3 Critères de convergence

Etant donné que les algorithmes peuvent être appliqués plusieurs fois de suite (cf section précédente), il nous faut définir un critère d'arrêt pour les différentes procédures d'estimation. Nous utilisons un critère classique de convergence basé sur la norme de la différence entre deux estimations successives de Θ_{fonc}. Ainsi, nous arrêtons l'algorithme après la K-ième itération complète si :

$$||\hat{\Theta}^{(K)}_{fonc} - \hat{\Theta}^{(K-1)}_{fonc}||^2 < \epsilon$$

avec $||.||$ la norme euclidienne sur $\mathbb{R}^{d_{\Theta_{fonc}}}$ et ϵ un seuil à définir (dans cette thèse, $\epsilon = 10^{-4}$ pour le filtre de Kalman et $\epsilon = 10^{-3}$ pour les méthodes particulaires plus difficiles à stabiliser). Afin de s'assurer de la convergence, nous pouvons améliorer le critère précédent en considérant plusieurs itérations de suite. Par exemple, nous arrêtons l'algorithme après la K-ième itération complète si :

$$\sum_{k=K-K_0+1}^{K} ||\hat{\Theta}^{(k)}_{fonc} - \hat{\Theta}^{(k-1)}_{fonc}||^2 < K_0\,\epsilon \tag{7.12}$$

avec K_0 à définir ($K_0 = 10$ dans cette thèse). Ceci permet de s'assurer que l'estimation de Θ_{fonc} se stabilise. Nous pouvons aussi imposer un rang minimum pour K pour éviter de tomber sur des minima locaux (rang que l'on peut définir avec l'indicateur de performance par exemple).

Concernant les méthodes particulaires, il est plus difficile de stabiliser les estimations $\hat{\Theta}^{(k)}_{fonc}$. En effet, ce sont des algorithmes stochastiques. Ils donneront donc une estimation différente à chaque itération complète de l'algorithme. Il existe une méthode appelée

« averaging » (voir Cappé et al. (2005)) qui permet de stabiliser l'estimation. Fixons une itération minimale K_0 et posons :

$$\forall K > K_0, \qquad \overline{\Theta}_{fonc}^{(K)} = \frac{1}{K - K_0} \sum_{k=K_0+1}^{K} \hat{\Theta}_{fonc}^{(k)}.$$

Le vecteur $\overline{\Theta}_{fonc}^{(K)}$ reste une estimation de Θ_{fonc} (voir Cappé et al. (2005)). Il est moins sujet aux fluctuations que $\hat{\Theta}_{fonc}^{(K)}$ ce qui permet de lui appliquer le critère de convergence donné par l'équation (7.12). Il est préférable de choisir un rang K_0 suffisamment élevé pour ne pas biaiser l'estimation de Θ_{fonc} avec de mauvaises estimations $\hat{\Theta}_{fonc}^{(K)}$. Les itérations 1 à K_0 sont alors appelées période de « burning ». Nous pouvons par exemple choisir un rang K_0 au-delà duquel le critère des moindres carrés est plus petit qu'un certain seuil κ :

$$K_0 = \min\{k,\ ||\overline{Y}(\hat{\Theta}_{fonc}^{(k)}) - Y_{obs}||^2 < \kappa\}$$

avec $||.||$ la norme euclidienne sur $\mathbb{R}^{d_{Y_{obs}}}$, $Y_{obs} = (Y_n)_{n\in\{0,\dots,T_{obs}\}}$ le vrai vecteur de données (regroupant tous les vecteurs d'observation Y_n) et $\overline{Y}(\Theta_{fonc}^{(k)})$ le vecteur de données calculé à partir du système dynamique avec $\hat{\Theta}_{fonc}^{(k)}$ en paramètres et sans pertubation (les bruits de modélisation et de mesure sont nuls).

N.B. 7.10 La définition de $\overline{\Theta}_{fonc}^{(K)}$ proposée ici est valable lorsque le nombre de particules ne change pas d'une itération complète de l'algorithme sur l'autre. Dans le cas contraire, d'autres définitions sont possibles, cf Cappé et al. (2005).

7.2.4 À propos du nombre de particules

Pour les méthodes de filtrage particulaire et filtrage particulaire convolé, il est possible (et même astucieux !) de changer le nombre de particules d'une itération complète de l'algorithme à l'autre, en particulier sur les premières itérations. Pour la première itération de l'algorithme, il est préférable de commencer avec un grand nombre de particules M_1 (cf section 7.2.1) afin d'orienter directement la recherche du vecteur de paramètres dans la bonne zone de l'espace des paramètres $\mathcal{P}$. En effet, plus M_1 est grand, plus les vecteurs initiaux $\tilde{\Theta}_0{}^{(i)}$ seront diversifiés et bien répartis dans $\mathcal{P}$ ce qui permet d'affiner d'emblée la recherche du vecteur optimal. En revanche, cela signifie un nombre de particules M plus conséquent ce qui ralentit considérablement l'algorithme. Etant donné que l'on cherche juste à placer globalement les particules dans la bonne zone de $\mathcal{P}$, il n'est pas nécessaire d'effectuer les $T_{obs} + 1$ étapes d'une itération complète de l'algorithme. Nous pouvons nous limiter aux cinq premières par exemple pour diminuer le temps de calcul. Une fois la première itération terminée, nous pouvons restreindre le nombre de particules et effectuer alors les $T_{obs} + 1$ étapes pour les itérations suivantes.

N.B. 7.11 Il est impératif d'utiliser les $T_{obs} + 1$ étapes pour les itérations $k \geq 2$ de l'algorithme. En effet, utiliser peu d'observations entraîne une perte d'information qui nuit à la qualité de l'estimation de Θ_{fonc}.

En résumé, il faut prendre beaucoup de particules bien réparties dans $\mathcal{P}$ pour la première itération mais en se limitant à un petit nombre d'étapes. Pour les itérations suivantes, on restreint le nombre de particules mais on utilise toutes les données. En faisant ainsi, on concentre d'abord les particules dans la bonne zone de $\mathcal{P}$ et ensuite on affine le résultat. Cette procédure permet d'accélérer grandement la convergence des algorithmes particulaires.

7.2.5 Modèle d'évolution pour le vecteur de paramètres, densité de transition d'importance et noyaux

• Modèle d'évolution pour le vecteur de paramètres
Dans la section 6.2.3, nous avons défini un modèle d'évolution pour le vecteur de paramètres Θ de la forme suivante :

$$\Theta_{n+1} = h_{n+1}(\Theta_n, U_{n+1}), \qquad n \geq 0. \tag{7.13}$$

Ce modèle permet de donner une dynamique d'évolution au vecteur d'état caché augmenté $X_n^a = (X_n, \Theta_n)$ et donc d'utiliser les méthodes d'inférence bayésienne. En ce qui concerne le filtre particulaire convolé, l'évolution du vecteur de paramètres se fait niveau de l'étape de correction lors du tirage des M particules $\{\tilde{x}_{n+1}^{a\,(i)},\, i = 1, \ldots, M\}$ suivant la densité estimée $\hat{p}(x_{n+1}^a | y_{0:n+1})$. Il n'est donc pas nécessaire de donner une dynamique d'évolution à Θ. Dans ce cas, $h_{n+1}(\Theta_n, U_{n+1}) = \Theta_n$. Par contre, il est nécessaire de le faire pour les trois autres algorithmes. Nous choisissons alors un modèle d'évolution de la forme :

$$\Theta_{n+1} = \Theta_n + U_{n+1}, \qquad n \geq 0$$

avec la contrainte que $\{U_n\}_{n\geq 0}$, $\{W_n\}_{n\geq 0}$ et $\{V_n\}_{n\geq 0}$ sont des suites mutuellement indépendantes de vecteurs aléatoires i.i.d.. Pour le filtre de Kalman sans parfum et le filtre particulaire de Rao-Blackwell, il est impératif d'imposer que U_n soit un vecteur gaussien. Par contre, ce n'est pas obligatoire pour le filtre particulaire. Dans cette thèse, nous avons choisi U_n vecteur gaussien centré de matrice de covariance L_n. Nous supposons L_n connue ou fixée dans un premier temps (voir la section 7.3.4 pour plus de détails). Dans ce cas, nous pouvons calculer la densité de transition $p(x_{n+1}^a | x_n^a)$ associée au modèle de Markov caché (respectivement $p(x_{n+1}^a | x_n^a, c_{n+1})$ pour les modèles à saut de Markov) :

$$\begin{aligned} p(x_{n+1}^a | x_n^a) &= p(x_{n+1}, \theta_{n+1} | x_n, \theta_n) \\ &= p(x_{n+1} | x_n, \theta_{n+1}) p(\theta_{n+1} | \theta_n) \\ &= \mathcal{N}\left(q'_{n+T_{exp}^m}; \phi_{n+T_{exp}^m}, K_{n+T_{exp}^m}\left(\phi_{n+T_{exp}^m}\right)^2\right) \mathcal{N}(\theta_{n+1}\,;\, \theta_n, L_n). \end{aligned} \tag{7.14}$$

avec :

$$\phi_{n+T_{exp}^m} = \Phi_{n+T_{exp}^m}(q_{(n+T_{exp}^m - T_{act})^+}, \ldots, q_{n+T_{exp}^m - 1}, N_{0\to n+T_{exp}^m}, E_{n+T_{exp}^m}, \Theta_{fonc})$$

et $x_n = (q_{n+T_{exp}^m - T}, \ldots, q_{n+T_{exp}^m - 1})$. Cette densité interviendra dans le calcul des poids (voir ci-dessous). Nous pouvons également calculer la matrice de covariance J_n associée

au vecteur d'état caché augmenté X_n^a (intervenant dans le filtre de Kalman sans parfum, cf la section 6.3.1) :

$$J_n = \left(\begin{array}{c|c} \mathbf{K}_n & \mathbf{0}_{d_{T,L_n}} \\ \hline \mathbf{0}_{d_{L_n,T}} & L_n \end{array} \right)$$

avec $\mathbf{0}_{d_1,d_2}$ la matrice de taille $d_1 \times d_2$ dont toutes les composantes sont nulles et $\mathbf{K}_n$ la matrice diagonale de taille $T \times T$ dont les éléments diagonaux sont donnés par le vecteur $(K_{n+T_{exp}^m-T}, \ldots, K_{n+T_{exp}^m-1})$.

• Densité de transition d'importance

Nous proposons quelques densités de transition d'importance $q_{n+1}(x_{n+1}^a|x_n^a)$ pour la méthode de filtrage particulaire. La densité la plus simple est la densité d'évolution du vecteur d'état caché augmenté :

$$q_{n+1}(x_{n+1}^a|x_n^a) = p(x_{n+1}^a|x_n^a)$$

Cette densité présente l'avantage d'être facilement simulable et de simplifier considérablement le calcul des poids donné par l'équation (6.19) du chapitre 6. La formule de récurrence liant les poids devient alors :

$$w_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)}) = w_n(\tilde{x}_{0:n}^{a\,(i)}) p(y_{n+1}|\tilde{x}_{n+1}^{a\,(i)}) = w_n(\tilde{x}_{0:n}^{a\,(i)})\, \mathcal{N}(y_{n+1}\,;\, g(\tilde{x}_{n+1}^{a\,(i)}), R_{n+1}).$$

Un gros inconvénient de cette densité est qu'elle ne tient pas compte de la nouvelle observation y_{n+1} pour choisir les M particules $\{\tilde{x}_{n+1}^{a\,(i)},\, i = 1, \ldots, M\}$. Les particules sont orientées sans information *a priori* ce qui peut entraîner une dégénérescence des poids et donc un rééchantillonnage plus fréquent. Une solution consiste alors à utiliser un pas de filtre de Kalman sans parfum. Pour chaque particule i, on calcule l'état corrigé $\hat{x}_{n+1|n+1}^a$ à partir de $\hat{x}_{n|n}^a = \tilde{x}_n^{a\,(i)}$ et de $\hat{\Sigma}_{n|n}^{x^a} = \mathbf{0}_{d_{X_n^a}, d_{X_n^a}}$ avec une étape de filtre de Kalman sans parfum. On pose alors $\tilde{x}_{n+1}^{a\,(i)} = \hat{x}_{n+1|n+1}^a$. Cette procédure a pour avantage de calculer $\tilde{x}_{n+1}^{a\,(i)}$ en fonction de y_{n+1} ce qui augmente l'efficacité du guidage des particules. En utilisant l'équation (7.14), l'équation de récurrence des poids devient dans ce cas :

$$\begin{aligned} w_{n+1}(\tilde{x}_{0:n+1}^{a\,(i)}) &= w_n(\tilde{x}_{0:n}^{a\,(i)})\, \frac{p(y_{n+1}|\tilde{x}_{n+1}^{a\,(i)}) p(\tilde{x}_{n+1}^{a\,(i)}|\tilde{x}_n^{a\,(i)})}{q_{n+1}(\tilde{x}_{n+1}^{a\,(i)}|\tilde{x}_n^{a\,(i)})} \\ &= w_n(\tilde{x}_{0:n}^{a\,(i)})\, \frac{\mathcal{N}(y_{n+1}\,;\, g(\tilde{x}_{n+1}^{a\,(i)}), R_{n+1})\, p(x_{n+1}^a|x_n^a)}{\mathcal{N}\left(q'_{n+T_{exp}^m}; \phi_{n+T_{exp}^m}, K_{n+T_{exp}^m}\left(\phi_{n+T_{exp}^m}\right)^2\right)} \\ &= w_n(\tilde{x}_{0:n}^{a\,(i)})\, \mathcal{N}(y_{n+1}\,;\, g(\tilde{x}_{n+1}^{a\,(i)}), R_{n+1})\, \mathcal{N}(\tilde{\theta}_{n+1}^{(i)}\,;\, \tilde{\theta}_n^{(i)}, L_{n+1}). \end{aligned}$$

avec :

$$\phi_{n+T_{exp}^m} = \Phi_{n+T_{exp}^m}(q_{(n+T_{exp}^m-T_{act})^+}, \ldots, q_{n+T_{exp}^m-1}, N_{0\to n+T_{exp}^m}, E_{n+T_{exp}^m}, \Theta_{fonc}).$$

Il existe d'autres façon de choisir la densité de transition d'importance (voir Liu and West (2001), Pitt and Shepard (2001) et Caron (2006)) : utilisation d'une variable auxiliaire, méthode de régularisation par noyau, distribution de Whishart.

• **Choix des noyaux** Nous définissons les noyaux K^X et K^Y (resp. K^C pour les modèles à saut de Markov) associés aux vecteurs aléatoires X_n^a et Y_n (resp. C_n). Pour K^X, on choisit la densité d'une loi normale centrée réduite. $K^X(x)$ est bien un noyau de Parzen-Rozenblatt puisque $|x|^{d_{X_n^a}} K^X(x) \to \infty$ quand $|x| \to \infty$. Introduisons alors le noyau réduit $K^X_{h_M^X}(x)$:

$$K^X_{h_M^X}(x) = \frac{1}{{h_M^X}^{d_{X_n^a}}} K^X(\frac{x}{h_M^X}) = \left(\frac{1}{\sqrt{2\pi} h_M^X}\right)^{d_{X_n^a}} e^{-|x|^2/{h_M^X}^2}.$$

avec $h_M^X > 0$ la fenêtre du noyau définie par :

$$h_M^X = C_X M^{-1/(4+d_{X_n^a})}.$$

C_X est une constante à ajuster (voir le *Nota Bene* 7.13 sur la réitération de l'algorithme). Nous choisissons un noyau réduit $K^Y_{h_M^Y}(y)$ de façon similaire pour Y_n en remplaçant h_M^X par $h_M^Y = C_Y M^{-1/(4+d_{Y_n})}$. Il a été prouvé que ce choix de h_M^X et de h_M^Y permet la convergence L^1 de $\hat{p}(x_n^a|y_{0:n})$ vers $p(x_n^a|y_{0:n})$ quand $M \to \infty$ (voir Campillo and Rossi (2009)).

N.B. 7.12 On peut améliorer les performances du filtre en choisissant astucieusement C_X à chaque itération de l'algorithme. Le paramètre C_X est lié à la dispersion des particules. Si C_X est petit, alors la particule ${\tilde{x}_{n+1}^a}^{(i)}$ sera proche de $\tilde{x}_n^{a\,(i)}$. Une grande valeur de C_X ($\approx 5.10^{-2}$) permet de bien explorer l'espace des états. Au contraire, un petit C_X ($< 10^{-3}$) concentre les particules dans une zone donnée. En tenant compte de ceci, l'idée est de prendre C_X grand pour les premières itérations afin de permettre une bonne exploration de l'espace des paramètres $\mathcal{P}$ par les particules. On affecte ensuite à C_X des valeurs de plus en plus petites au fur et à mesure des itérations afin de réduire la variance des particules et améliorer la précision de l'estimation du vecteur de paramètres Θ. La constante C_Y reste la même tout au long de la méthode. Elle est fixée de façon à donner aux poids des particules un ordre de grandeur raisonnable (il faut éviter de n'avoir que des poids très petits car cela entraîne des instabilités numériques).

En ce qui concerne le noyau K^C, il n'est pas très judicieux de choisir un noyau gaussien car les composantes de C_n sont des entiers naturels. De plus, C_n a une signification botanique très précise puisqu'il représente l'évolution de la composition de la plante. Il n'est donc pas raisonnable de le perturber aléatoirement comme pour X_n et Y_n (les nouvelles plantes que l'on obtiendrait après perturbations peuvent n'avoir aucun sens dans la réalité). Pour ces raisons, on choisit de ne pas perturber C_n. Son noyau K^C est donc donné par :

$$K^C_{h_M^C}(c) = \delta_0(c).$$

7.3 Analyse et comparaison des méthodes

L'objectif de cette section est d'analyser les performances des méthodes d'inférence bayésiennes pour l'estimation du vecteur de paramètres Θ_{fonc} et de les comparer entre

elles. Dans cette optique, elles sont appliquées sur trois cas-tests utilisant le modèle de croissance de plante GreenLab (voir la section 2.3 pour une description complète du modèle) ayant les spécificités suivantes :

- Cas-test 1 (= cas-test simple) : évolution de la structure de la plante déterministe avec temps d'expansion et d'activité égaux à 1 ;
- Cas-test 2 (cas de la betterave, voir les sections 7.3.5 et 7.5) : évolution de la structure de la plante déterministe avec temps d'expansion et d'activité supérieurs à 1 ;
- Cas-test 3 (cf section 7.3.6) : évolution de la structure de la plante stochastique avec temps d'expansion et d'activité égaux à 1.

Les cas-tests sont réalisés à partir de données simulées ce qui nous permet de connaître le vrai vecteur de paramètres Θ_{fonc} noté Θ_{fonc}^{vrai} et de bien comprendre le comportement des méthodes. Tout comme dans la section précédente, nous supposons connues ou fixées la biomasse Q_0 contenue dans la graine ainsi que les variances K_n et matrices de covariance R_n, $n \in \mathbb{N}$, associées respectivement aux bruits de modélisation et aux bruits de mesure (voir la section 7.3.4 concernant l'impact de cette hypothèse sur l'estimation). Dans cette section, nous nous intéressons uniquement à l'estimation du vecteur de paramètres Θ_{fonc} et non à sa distribution *a posteriori*. La section 7.4 est dédiée à ce sujet.

L'objectif du premier cas-test est d'analyser les performances du filtre de Kalman sans parfum, du filtre particulaire et du filtre particulaire convolé dans le cadre d'un modèle de croissance simple (peu de paramètres à estimer) et faiblement non-linéaire. Nous commençons tout d'abord par décrire ce cas-test dans la section 7.3.1. Ensuite, divers aspects des méthodes sont étudiés (sections 7.3.2, 7.3.3 et 7.3.4). Le second cas-test (cf section 7.3.5) a pour but d'analyser le passage à des temps d'expansion et d'activité élevés augmentant la non-linéarité du système. Pour le dernier cas-test (section 7.3.6), nous cherchons à comparer les méthodes de filtrage particulaire convolé et filtrage particulaire de Rao-Blackwell dans le cas où l'évolution de la structure de la plante est stochastique. Pour les deux derniers cas-tests, il n'est pas nécessaire de faire l'étude complète des méthodes comme pour le premier cas-test car la plupart des résultats obtenus restent valables. Nous terminons par un bilan des différentes méthodes en soulignant leurs points forts et leurs points faibles, cf section 7.3.7.

N.B. 7.13 Tous les cas-tests ont été implémentés avec Matlab 7.9.0 (R2009b) et les simulations de données en C++.

7.3.1 Description du cas-test simple

Considérons un modèle de type GreenLab 1 dont le modèle de développement $\mathcal{S}$ est le suivant :

- La classe physiologique maximale est $CP_m = 2$;
- Un entrenœud de CP 1 porte un bourgeon apical de CP 1, un bourgeon axillaire de CP 2 et une feuille ;

– Un entrenœud de CP 2 porte un bourgeon apical de CP 2 et une feuille ;
– Les bourgeons ont un comportement déterministe et produisent toujours un nouvel entrenœud à chaque cycle de développement.

Les règles de production de $\mathcal{S}$ sont illustrées par la figure 7.1.

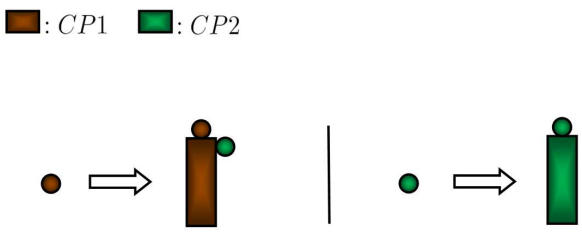

FIG. 7.1 – Règles de production pour le cas-test simple. CP = Classe Physiologique.

Les ensembles $\mathcal{B}$ et $\mathcal{O}$ associés aux bourgeons et aux organes sont donnés par :

$$\mathcal{B} = \{b_1, b_2\} \qquad \mathcal{O} = \{e_1, e_2, l, p\}.$$

Dans ce cas-test, les feuilles ont un temps d'activité de 1 et les organes un temps d'expansion de 1 :

$$T_{act} = 1 \qquad \forall o \in \mathcal{O}, \quad T^o_{exp} = 1.$$

Ainsi, les puits des différents organes sont constants et donnés par un seul paramètre P_o avec $o \in \mathcal{O}$. L'ensemble des paramètres du modèle est donc donné par le tableau 7.1. L'objectif est de réaliser l'estimation du vecteur de paramètres Θ_{fonc} défini par :

$$\Theta_{fonc} = (\mu, S_p, P_p, P_{e_1}, P_{e_2})\,.$$

Afin de tester le comportement des méthodes, nous avons également fait varier le nombre de paramètres à estimer (de 1 à 5). Dans ce cas, le vecteur à estimer est un sous-vecteur de Θ_{fonc}. Les autres paramètres de θ_{fonc} sont alors fixés à leur vraie valeur. L'estimation se fait à partir d'une plante virtuelle mesurée à son vingt-huitième cycle de développement. Nous disposons donc d'un jeu de données $\{y_k,\ k = 0, \ldots, T_{obs}\}$ avec $T_{obs} = 28$ (voir la section 7.1.1 pour une description complète des données botaniques). La figure 7.2 donne le comportement du système dynamique au travers des 28 cycles de développement.

7.3.2 Stabilité de l'estimation et biais

Nous cherchons à comparer la qualité des estimations du filtre de Kalman sans parfum, du filtre particulaire et du filtre particulaire convolé pour le cas-test simple. Les estimations sont obtenues après convergence des algorithmes selon le critère défini dans la section 7.2.3. Auparavant, il faut s'assurer de la stabilité des estimations fournies par les différents algorithmes. Comme l'algorithme du filtre de Kalman sans parfum est déterministe, celui-ci fournit donc toujours la même estimation des paramètres pour un même jeu de données $y_{0:T_{obs}}$. En revanche, ce n'est pas le cas pour les deux autres

Paramètres	Description	Valeur	Nature	Unité
$E(n)$	Facteur environnement au cycle n	2000	Mesuré	$J.m^{-2}$
K_n	Variance du bruit de modélisation	6.25×10^{-4}	Fixé	Sans Unité
σ_o	Ecart type du bruit de mesure	2.5×10^{-2}	Fixé	g
μ	Water use efficiency	0.095	Estimé	$g.J^{-1}$
S_p	Surface foliaire spécifique	2.5	Estimé	m^2
k_b	Coefficient d'extinction de la loi de Beer-Lambert	2.5×10^{-2}	Fixé	Sans Unité
P_l	Force de puits d'un limbe	1	Fixé	Sans Unité
P_p	Force de puits d'un pétiole	0.3	Estimé	Sans Unité
P_{e_1}	Force de puits d'un entrenœud de classe physiologique 1	3	Estimé	Sans Unité
P_{e_2}	Force de puits d'un entrenœud de classe physiologique 2	2	Estimé	Sans Unité
e	Masse surfacique d'une feuille	10	Mesuré	$g.m^{-2}$
Q_0	Biomasse contenue dans la graine	10	Mesuré	g

TAB. 7.1 – Valeurs des paramètres utilisés pour générer les données. Le paramètre S_p rend compte de l'effet compétition. « Mesuré » signifie que le paramètre est mesuré directement à partir de données botaniques. « Fixé » signifie que la valeur des paramètres a été fixée pour des raisons d'identifiabilité. Enfin, « Estimé » signifie qu'il s'agit d'un paramètre à estimer à partir des données. Dans ce cas, le chiffre indiqué dans la case valeur correspond à la vraie valeur du paramètre (c'est celle qui a servi à créer le jeu d'observations simulé). Nous avons donc $\Theta_{fonc}^{vrai} = (0.095, 2.5, 0.3, 3, 2)$.

méthodes particulaires. Pour chacune d'entre elles, une estimation différente de Θ_{fonc} est obtenue après chaque exécution de l'algorithme. Pour tester la variabilité des estimations avec un jeu d'observations $y_{0:T_{obs}}$ donné, les algorithmes de filtrage particulaire simple et convolé ont été exécutés 200 fois. Ainsi, 200 vecteurs de paramètres estimés ont été obtenus pour les deux méthodes. Leur distribution est représentée par la figure 7.3.

Globalement, les distributions sont centrées autour des mêmes valeurs pour chacun des paramètres. En revanche, il apparaît clairement que la variabilité des estimation est plus importante pour le filtrage particulaire simple que pour le convolé. Ce dernier est même relativement précis dans ces estimations. Le tableau 7.2 propose de comparer les valeurs moyennes obtenues par les trois algorithmes pour le même jeu de données $y_{0:T_{obs}}$.

Les trois méthodes proposent des estimations très proches l'une de l'autre. Notons tout de même la présence d'un biais concernant l'estimation de μ et de S_p qui est dû au peu de données disponibles (29 vecteurs d'observation uniquement). Les estimations des puits des organes sont très bonnes (particulièrement celles du filtre de Kalman). Ce tableau confirme à nouveau que la variance des estimations est plus importante pour le filtre particulaire simple que pour le convolé.

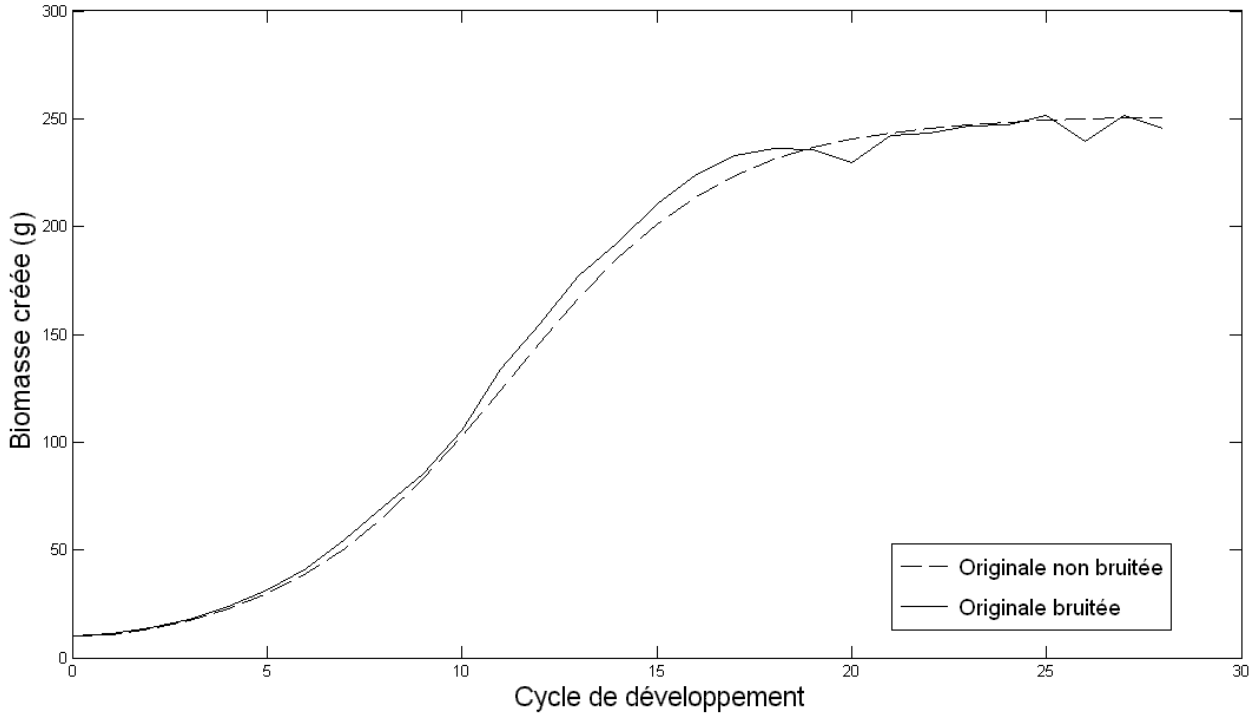

FIG. 7.2 – Biomasse créée par photosynthèse à chaque cycle de développement à partir du vrai vecteur de paramètres Θ_{fonc}^{vrai} : la courbe en trait plein représente la vraie valeur de la biomasse créée pour une réalisation du système dynamique (incluant le bruit de modélisation). Celle en traits pointillés représente la biomasse obtenue par une équation de photosynthèse sans bruit.

Paramètres	Vraie valeur	Est. FK	Est. FP	Est. FPC	Std FP	Std FPC
μ	9.5×10^{-2}	9.230×10^{-2}	9.257×10^{-2}	9.285×10^{-2}	2.64×10^{-3}	2.14×10^{-3}
S_p	2.5	2.5974	2.5921	2.5834	8.56×10^{-2}	7.08×10^{-2}
P_p	0.3	0.3000	0.2987	0.2992	6.51×10^{-3}	4.03×10^{-3}
P_{e_1}	3	3.0006	2.9949	2.9967	1.59×10^{-2}	1.18×10^{-2}
P_{e_2}	2	1.9923	2.0092	1.9974	1.06×10^{-2}	7.62×10^{-3}

TAB. 7.2 – Estimation des paramètres et écart-type des estimations. Est. = Estimation. Std = écart-type. FK = Filtre de Kalman sans parfum. FP = Filtre Particulaire. FPC = Filtre Particulaire Convolé.

Pour comparer la qualité des estimations des différentes méthodes, nous avons généré 200 jeux de données $y_{0:T_{obs}}^{(i)}$, $i \in \{1, \ldots, 200\}$, à partir du même vecteur de paramètres Θ_{fonc}^{vrai} (voir le tableau 7.1). Pour chacun de ces jeux, les trois algorithmes ont été exécutés une seule fois et les vecteurs de paramètres estimés recueillis. La valeur moyenne de chacun des paramètres estimés ainsi que leur écart-type est regroupé dans le tableau 7.3.

Les estimations sont globalement très bonnes. Notons tout de même la présence d'un biais sur les paramètres de l'équation de production μ et S_p. Le filtre particulaire propose

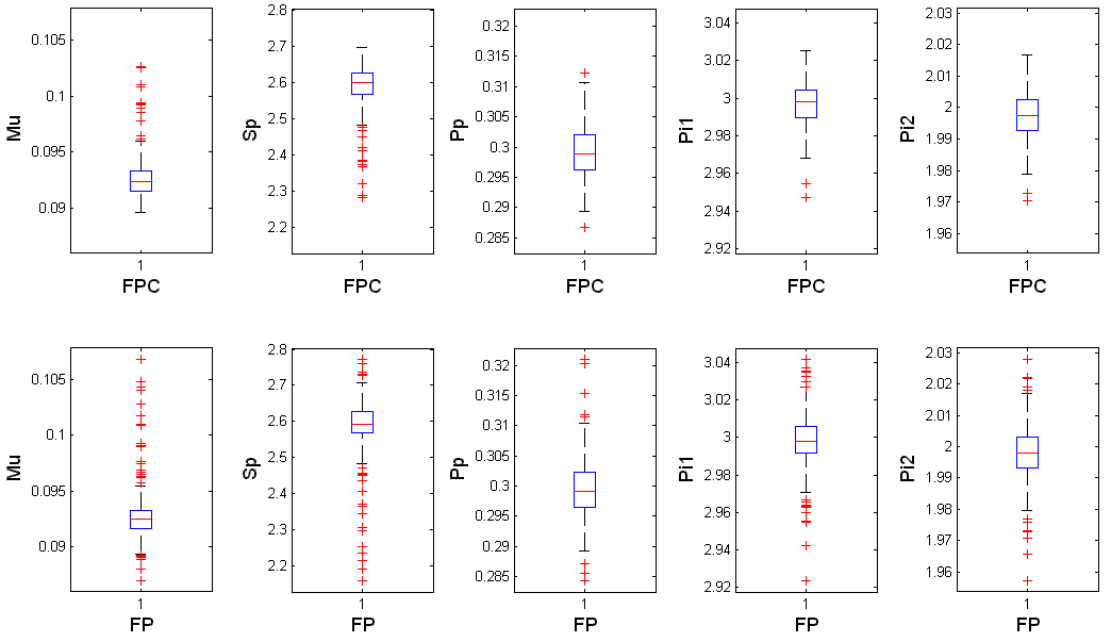

FIG. 7.3 – Comparaison des distributions des paramètres pour les méthodes de filtrage particulaire (= FP) et filtrage particulaire convolé (= FPC). Le trait rouge indique la médiane, les bleus indiquent les quartiles et les noirs les valeurs extrêmes. Les croix correspondent aux mesures abérantes (outliers). On considère qu'une mesure est abérante si elle est située à plus de 1.5 fois l'écart interquartile à partir des limites de la boîte.

Paramètres	Vraie valeur	Est. FK	Est. FP	Est. FPC	Std FK	Std FP	Std FPC
μ	9.5×10^{-2}	9.436×10^{-2}	9.456×10^{-2}	9.434×10^{-2}	3.59×10^{-3}	4.39×10^{-3}	2.96×10^{-3}
S_p	2.5	2.5292	2.5158	2.5230	1.26×10^{-1}	1.49×10^{-1}	9.71×10^{-2}
P_p	0.3	0.3000	0.2994	0.2997	4.50×10^{-4}	8.59×10^{-3}	6.08×10^{-3}
P_{e_1}	3	3.0003	2.9982	2.9984	2.4×10^{-3}	2.21×10^{-2}	1.58×10^{-2}
P_{e_2}	2	2.0002	1.9983	1.9985	1.35×10^{-3}	1.49×10^{-2}	1.09×10^{-2}

TAB. 7.3 – Estimation moyenne des paramètres et écart-type des estimations pour 200 jeux de données générés avec le même vecteur de paramètres Θ_{fonc}^{vrai}. Est. = Estimation. Std = écart-type. FK = Filtre de Kalman sans parfum. FP = Filtre Particulaire. FPC = Filtre Particulaire Convolé.

les meilleures estimations de μ et S_p avec le plus petit écart-type. Le filtre de Kalman propose d'excellentes estimations pour les puits avec un écart-type très faible. Le filtre particulaire reste le moins performant des trois sur tous les paramètres que ce soit au niveau de l'estimation ou de son écart-type. Il est raisonnable d'appliquer une seule fois les algorithmes pour le FPC à cause de la faible variabilité des estimations ce qui est moins le cas pour le filtre particulaire. Notons également que les méthodes sont toutes très robustes (pour chaque jeu de données $y_{0:T_{obs}}^{(i)}$, $i \in \{1, \ldots, 200\}$, les estimations du vecteur de paramètres restent très proches de Θ_{fonc}^{vrai} avec un faible écart-type).

Nous pouvons également définir un critère de performance pour comparer les méthodes (à première vue, il n'est pas facile de comparer le filtre de Kalman et le filtre particulaire convolé). Un choix possible est l'écart moyen à la vraie valeur des para-

mètres. Pour $i \in \{1, \ldots, 200\}$, notons provisoirement $\hat{\Theta}_m^{(i)}$ l'estimation obtenue par la méthode m à partir du jeu de données $y_{0:T_{obs}}^{(i)}$. Le critère de performance $\mathscr{C}_m$ associé à cette méthode m vaut :

$$\mathscr{C}_m = \frac{1}{200} \sum_{i=1}^{200} ||\hat{\Theta}_m^{(i)} - \Theta_{fonc}^{vrai}||^2 \tag{7.15}$$

avec $||.||$ la norme euclidienne sur $\mathbb{R}^{d_{\Theta_{fonc}}}$. La meilleure méthode est donc celle qui présente le plus petit $\mathscr{C}_m$. Dans le cas où les paramètres ont des ordres de grandeur différents, il est préférable de travailler avec des données centrées-réduites (nous divisons alors chaque composante du vecteur $\hat{\Theta}_m^{(i)} - \Theta_{fonc}^{vrai}$ dans l'équation (7.15) par la composante de Θ_{fonc}^{vrai} correspondante). Le critère ainsi calculé est nommé critère réduit et se note $\mathscr{C}_m^r$. Le tableau 7.4 donne les critères de performance simples et réduits pour chacune des méthodes.

	Filtre Kalman	Filtre Particulaire	Filtre Particulaire Convolé
Critère $\mathscr{C}_m$	1.66×10^{-2}	4.63×10^{-2}	1.03×10^{-2}
Critère réduit $\mathscr{C}_m^r$	8.515	8.540	8.525

Tab. 7.4 – Critère de performance standard et réduit pour chacune des méthodes.

Le filtre de Kalman apparaît comme le plus performant des trois algorithmes avec le critère réduit mais pas avec le critère standard, le meilleur étant dans ce cas le filtre particulaire convolé. Les différences sont cependant minimes. Il est tout de même préférable de travailler avec le critère réduit afin de donner un poids équivalent à chacune des composantes du vecteur de paramètres (et ne pas privilégier les composantes à valeurs importantes). Le tableau confirme également les moins bonnes performances du filtre particulaire simple.

Les trois méthodes donnent également des résultats très bons concernant l'estimation des états cachés du système dynamique (c'est-à-dire la biomasse créée à chaque cycle de croissance dont la valeur est donnée par l'équation de photosynthèse bruitée (7.7)). Ceci est bien mis en évidence par la figure 7.4. Il est à noter que le filtre de Kalman sans parfum propose une estimation très précise de ces états cachés (voir la figure 7.5).

.

7.3.3 Convergence et vitesse de convergence

Dans cette section, les questions relatives à la convergence des algorithmes sont traitées : nombre d'itérations avant convergence, temps d'exécution des algorithmes, facteurs influençant la convergence... Pour ce faire, les méthodes ont été appliquées sur le cas-test simple pour un même jeu de données. Pour chacune des méthodes, nous avons également fait varier le nombre de paramètres à estimer pour observer son effet sur la convergence. Les résultats de ces tests sont résumés dans le tableau 7.5.

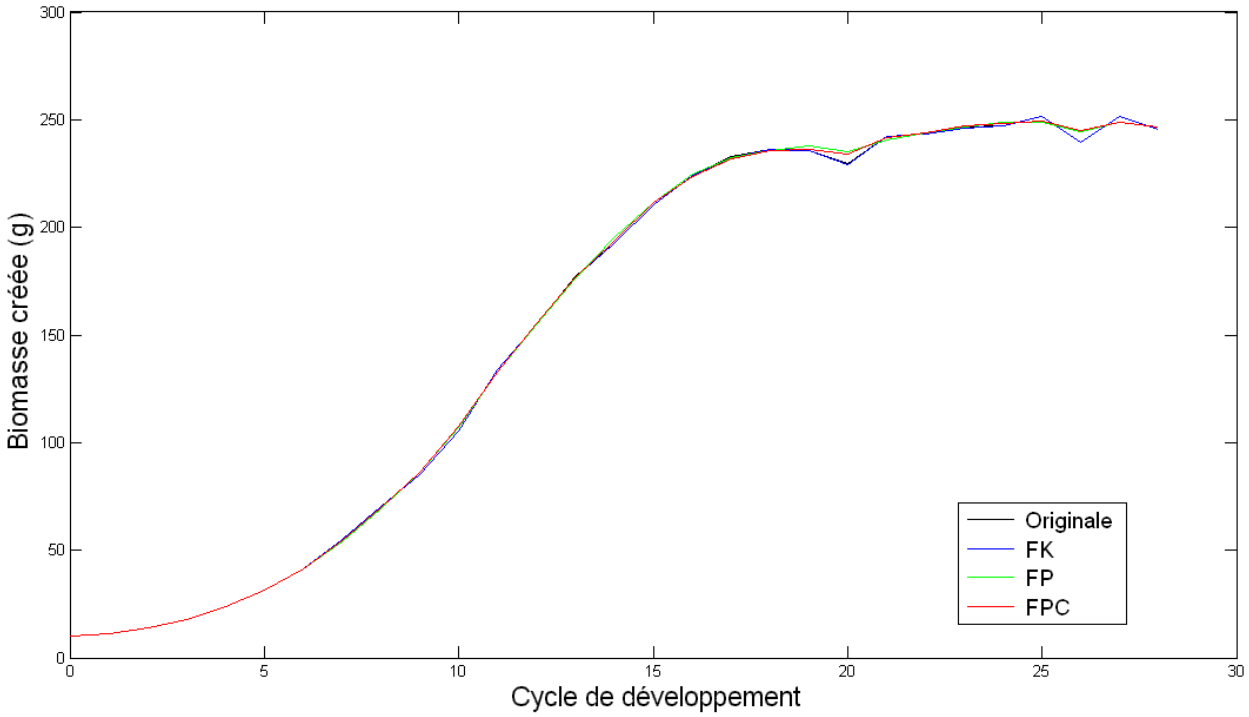

FIG. 7.4 – Estimation des états cachés du cas-test simple. Les courbes originale et UKF d'une part et FP et CPF d'autre part sont pratiquement confondues. FK = Filtre de Kalman sans parfum, FP = Filtre Particulaire, FPC = Filtre Particulaire Convolé.

	1 paramètre			3 paramètres			5 paramètres		
Méthode utilisé	FK	FP	FPC	FK	FP	FPC	FK	FP	FPC
Temps d'exécution (min)	2.3	1.3	1.5	10.8	11.4	12.2	24.7	25.2	28.2
Nombre d'itérations	2180	32	30	6348	53	49	12732	59	56
Nombre de particules	/	300	300	/	1500	1500	/	3000	3000

TAB. 7.5 – Comparaison des temps d'exécution et du nombre d'itérations avant convergence. FK = Filtre de Kalman. FP = Filtre Particulaire. FPC = Filtre Particulaire Convolé.

N.B. 7.14 Bien que les exemples proposés ne reposent que sur un seul jeu de données $y_{0:N}$, les résultats restent vrais de façon générale.

Il est bien entendu évident que le nombre d'itérations et en conséquence le temps d'exécution augmentent avec le nombre de paramètres à estimer. Nous pouvons remarquer que les temps d'exécution des trois algorithmes sont sensiblement du même ordre de grandeur. Le filtre de Kalman est un peu plus lent quand le nombre de paramètres à estimer est plus faible mais devient le plus rapide quand 3 paramètres ou plus sont à estimer. Notons également que le filtre particulaire simple est un peu plus rapide que dans sa version convolée. Cela est dû à la procédure de rééchantillonnage qui est systématique à chaque étape du filtre convolé.

Un point qui peut surprendre est le nombre d'itérations nécessaires à la convergence du filtre de Kalman. Une itération complète du filtre n'apporte en fait qu'une faible

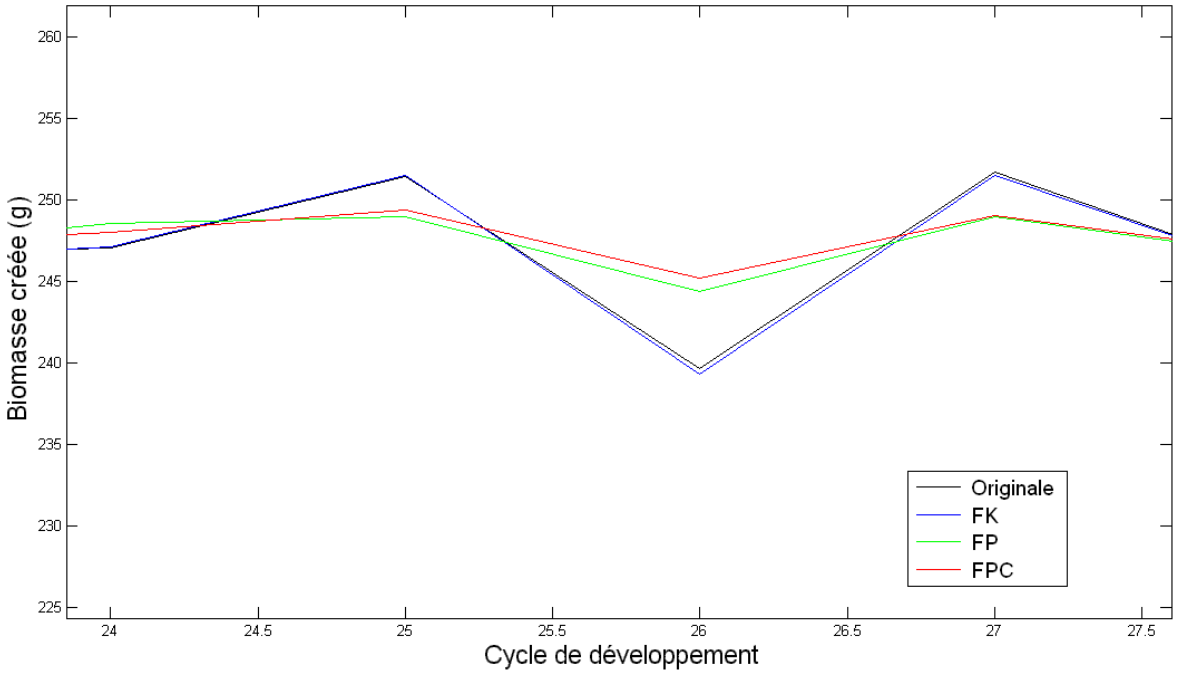

FIG. 7.5 – Agrandissement du cadre de la figure 7.4. La courbe UKF est encore pratiquement confondue avec l'originale ce qui montre la précision de l'estimation des états cachés. FK = Filtre de Kalman sans parfum, FP = Filtre Particulaire, FPC = Filtre Particulaire Convolé.

correction au vecteur de paramètres Θ_{fonc} surtout sur les dernières itérations. En fait, le filtre aboutit rapidement à un vecteur de paramètres estimé correcte en seulement 10 pourcents des itérations. Au-delà, les corrections apportées sont mineures. En résumé, le filtre trouve rapidement un vecteur de paramètres du bon ordre de grandeur mais il a besoin de beaucoup d'itérations pour affiner le résultat. Celui-ci se ratrappe bien évidemment par le faible temps d'exécution d'une itération complète de l'algorithme. Les méthodes particulaires ont un comportement inverse. Il faut beaucoup moins d'itérations pour vérifier le critère de convergence. Le nombre d'itérations augmente avec le nombre de paramètres à estimer mais cette augmentation est bien moins importante que dans le cas du filtre de Kalman. En revanche, le temps d'exécution d'une itération augmente radicalement avec la taille du vecteur de paramètres Θ_{fonc} à estimer. En effet, le temps d'exécution d'une itération dépend du nombre de particules utilisé. Plus le nombre de paramètres à estimer est important, plus la taille de l'espace des états augmentés $\mathcal{X}^a$ l'est aussi. Il faut donc beaucoup de particules pour pouvoir l'explorer correctement. Si le nombre de particules est faible, alors l'espace des paramètres $\mathcal{P}$ est mal exploré. Le jeu de particules obtenus à la fin de l'étape T_{obs} n'est pas vraiment représentatif des différentes évolutions possibles du système dynamique ce qui conduit à une estimation des paramètres du modèle non seulement erronée mais en plus instable. En effet, un faible nombre de particules ne permet pas la convergence du filtre. La variabilité des particules est alors telle que l'on obtient une estimation différente à chaque itération de l'algorithme. Il est donc nécessaire de prendre un nombre suffisant de particules (voir la section 7.3.7). D'ailleurs, il s'agit ici d'une des principales limites de ces méthodes

particulaires puisque si le nombre de paramètres à estimer est très grand, alors il faut une puissance de calcul et des capacités de stockage importantes pour pouvoir les utiliser.

En résumé, les trois méthodes convergent bien si les conditions suivantes sont vérifiées :

- Filtre de Kalman sans parfum : les conditions initiales sont du bon ordre de grandeur et les bruits de modélisation et de mesure pas trop importants (voir la section 7.3.4) ;
- Méthodes particulaires : le nombre de particules M doit être suffisamment important.

La vitesse de convergence des algorithmes dépend fortement du nombre de paramètres à estimer. Pour un nombre de paramètres supérieur à 3, le filtre de Kalman devient vite avantageux. En revanche, toutes ces méthodes montrent leurs limites si le nombre de paramètres devient trop important. De façon générale, il s'agit même d'une limitation qui est propre aux méthodes d'inférence bayésienne.

7.3.4 Influence des conditions initiales et des bruits sur l'estimation

Dans les sections 7.3.2 et 7.3.3, nous avons supposé connues ou fixées la biomasse Q_0 contenue dans la graine ainsi que les variances K_n et matrices de covariance R_n, $n \in \mathbb{N}$, associées respectivement aux bruits de modélisation et aux bruits de mesure. Dans cette section, nous traitons de leur influence sur l'estimation.

Influence des conditions initiales

Afin de permettre la convergence des algorithmes, il est important d'attribuer aux vecteurs d'état cachés initiaux des valeurs correspondant à l'ordre de grandeur des quantités qu'il représente. Dans le cas du filtre de Kalman sans parfum, un choix de x_0^a très éloigné de sa vraie valeur peut aboutir à une divergence du filtre. En effet, dans ce cas là, les observations prédites par les équations du filtre sont elles aussi très éloignées des vraies observations du système. Cela engendre une correction du vecteur d'état caché très importante pouvant aboutir à des valeurs négatives pour certains paramètres qui, normalement, sont positifs. Ceci est dû à l'hypothèse gaussienne faite sur la distribution de X_n^a conditionnellement à $Y_{0:n}$. En effet, cette hypothèse autorise des valeurs négatives pour X_n^a même si celles-ci n'ont aucun sens du point de vue physique pour le système dynamique ce qui le fait diverger. Pour les méthodes particulaires, une initialisation inappropriée pointe les particules dans la mauvaise zone de l'espace d'état caché $\mathcal{X}^a$ ce qui entraîne des poids très faibles pour les particules et la divergence des algorithmes. Concernant le cas-test simple, l'ordre de grandeur des différents paramètres est donné par le tableau 7.6.

Si la biomasse initiale Q_0 n'est pas connue, il est possible de l'estimer en l'incluant dans le vecteur des paramètres Θ_{fonc}. Il suffit alors de connaître un ordre de grandeur raisonnable (par exemple entre 3 et 25 pour le cas-test simple). En utilisant le filtre de

	μ	S_p	P_p	P_{e_1}	P_{e_2}	Q_0
Vraie valeur	0.095	2.5	0.3	3	2	10
Ordre de grandeur	$[0.01; 0.5]$	$[1; 5]$	$[0.1; 1]$	$[1; 10]$	$[1; 10]$	$[3; 25]$

TAB. 7.6 – Ordre de grandeur des paramètres du cas-test simple. Les ordres de grandeur peuvent être obtenus en utilisant l'interprétation botanique des paramètres. Il est possible également d'estimer au préalable les paramètres grâce à des estimateurs classiques comme les moindres carrés généralisés ce qui nous donne une idée de leur valeur.

Kalman avec 6 paramètres à estimer (les 5 initiaux de Θ_{fonc} et Q_0), Q_0 est estimé à 9.91 pour le jeu de données ayant servi à la création du tableau 7.5 (contre 10 pour le vrai Q_0). Ce léger biais est dû encore une fois au faible nombre de vecteurs d'observation.

Influence des bruits de modélisation et de mesure

Les bruits de modélisation et de mesure influencent peu l'estimation des paramètres du modèle. Dans le cas des méthodes particulaires, des bruits trop grands engendreraient une variance importante au niveau des particules et donc une variance importante pour l'estimation de Θ_{fonc} ce qui peut nuire à la convergence des algorithmes. En revanche, des bruits trop faibles diminueraient considérablement la vitesse de convergence. Il vaut donc tout de même mieux privilégier des petits bruits. Les estimations obtenues sont très proches quelque soit la valeur accordée aux bruits en restant dans une gamme raisonnable (valeurs diagonales des matrices de covariance inférieures à 10^{-2}). Concernant le filtre de Kalman sans parfum, le choix des bruits influencent peu l'estimation des paramètres (voir le tableau 7.7).

Paramètres	Vraie valeur	Estimation avec $K_n = 6.25 \times 10^{-4}$ et $\sigma_o = 2.5 \times 10^{-2}$	Estimation avec $K_n = 10^{-2}$ et $\sigma_o = 10^{-1}$
μ	9.5×10^{-2}	9.230×10^{-2}	9.102×10^{-2}
S_p	2.5	2.5971	2.6542
P_p	0.3	0.3001	0.3004
P_{e_1}	3	3.0006	3.0051
P_{e_2}	2	1.9992	2.0020

TAB. 7.7 – Influence des bruits de modélisation et de mesure sur l'estimation du filtre de Kalman sans parfum.

Cependant, si les bruits sont trop grands, le filtre peut diverger. En effet, l'hypothèse consistant à approcher $\hat{p}(x_n^a|y_{0:n})$ par une loi normale n'est plus raisonnable et les équations du filtre ne sont plus valables. Il est donc préférable également d'utiliser des petites valeurs de bruit même si la convergence est plus longue à obtenir. C'est pour ces raisons que les éléments diagonaux de la matrice de covariance L_n associée au vecteur gaussien U_n (voir l'équation d'évolution (7.13) associée à Θ_{fonc}) sont choisis égaux à 10^{-4}. Il en est de même pour la matrice B correspondant au bruit des paramètres dans l'initialisation de Σ_0^x pour le filtre de Kalman sans parfum.

N.B. 7.15 Rappelons que nous ne nous intéressons ici qu'à l'estimation du vecteur de paramètres Θ_{fonc}. Le choix des bruits de modélisation et de mesure a bien entendu une grande influence concernant l'estimation de la densité *a posteriori* des paramètres. Cependant, il est bon de noter que, sous certaines conditions, l'estimation du vecteur d'état caché x_n^a, $n \in \{0, \ldots, T_{obs}\}$, n'est pas trop influencée par le choix des bruits. Ceci nous permettra de proposer une procédure d'estimation des bruits dans la section 7.4 et d'en déduire la distribution *a posteriori* du vecteur de paramètres Θ_{fonc} par bootstrap paramétrique.

7.3.5 Comportement avec des temps d'activité et d'expansion quelconques

Le comportement des trois méthodes est étudié lorsque les temps d'activité et d'expansion ne sont plus égaux à 1. Le cas-test choisi est celui de la betterave. La description du modèle ainsi que les résultats de l'estimation sont présentés dans la section 7.5. Pour ce cas-test, nous avons travaillé avec des données réelles et non simulées. Nous ne donnons ici que les principaux résultats.

En ce qui concerne la betterave, les temps d'expansion et d'activité sont élevés. La non-linéarité du système est plus importante. Le filtre de Kalman devient donc inutilisable car divergent. Etant donné que la méthode de filtrage particulaire simple est plus instable que sa version convolée, seule cette dernière a été appliquée. Les estimations obtenues sont moins stables que pour le cas-test simple (voir la section 7.3.2). Ceci est dû au nombre de particules limité que nous avons pu utiliser. Nous n'avons pu prendre que 8000 particules (ce qui correspondait à la limite de la mémoire accordée par Matlab 7.9.0 (R2009b)) alors que 12000 auraient été plus appropriées. En effet, comme les temps d'expansion sont élevés, il est nécessaire de simuler des trajectoires sur un nombre important de cycles avant de pouvoir utiliser la première observation du système. Les particules « voyagent » donc dans l'espace des états sans guidage pendant plusieurs étapes de suite. Ceci entraîne une mauvaise exploration de l'espace en question si le nombre de particules est insuffisant. L'algorithme s'initialise donc très mal ce qui explique le manque de stabilité des résultats.

Les valeurs des paramètres estimés restent bonnes au vu des données réelles. Le temps d'exécution du filtre est de deux jours en moyenne sur un ordinateur standard et une soixantaine d'itérations complètes sont nécessaires pour vérifier le critère de convergence. Il apparaît donc clairement que la complexité du système dynamique augmente de façon drastique les temps d'exécution et nombres d'itérations.

7.3.6 Cas où l'évolution de la structure est stochastique

Nous reprenons le cas-test simple de la section 7.3.1 mais cette fois en considérant un modèle de développement $\mathcal{S}$ stochastique :

– un bourgeon de CP 1 meurt avec une probabilité $1 - p_s(1)$ ou bien il produit un phytomère de CP 1 portant un bourgeon apical de CP 1, un bourgeon latéral de CP 2 et une feuille avec une probabilité $p_s(1)$.

– un bourgeon de CP 2 produit toujours un phytomère de CP 2 portant un bourgeon apical de CP 2 et une feuille.

Les règles de production de $\mathcal{S}$ sont données par la figure 7.6.

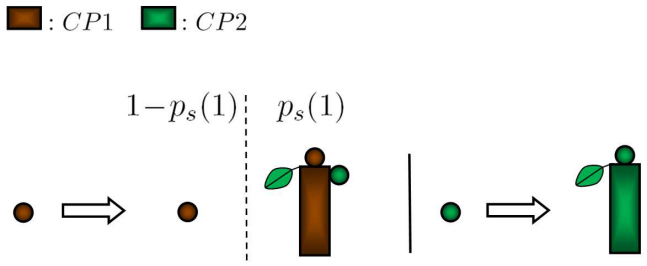

FIG. 7.6 – Règles de production associées au cas-test stochastique. CP = Classe Physiologique.

Les paramètres du modèle sont toujours donnés par le tableau 7.1 et le vecteur Θ_{fonc} par :

$$\Theta_{fonc} = (\mu, S_p, P_p, P_{e_1}, P_{e_2})\,.$$

Si l'évolution de la structure de la plante donnée par $\{N_n\}_{n\geq 0}$ est déterministe, alors les méthodes de filtrage de kalman sans parfum, de filtrage particulaire simple et convolé peuvent être utilisées pour l'estimation du vecteur de paramètres Θ_{fonc}. Dans cette section, nous supposons que cette évolution est stochastique. Le modèle de croissance $\mathcal{M}$ est alors un modèle à saut de Markov, cf théorème 7.1.2. Les méthodes de filtre particulaire simple, convolé et Rao-Blackwellisé peuvent donc être employées. Le filtre de Kalman ne peut être considéré car l'approximation de la loi de (X_n^a, C_n) conditionnellement à $Y_{0:n}$ par une gaussienne n'est pas raisonnable (le processus $\{C_n\}_{n\geq 0}$ a la même dynamique d'évolution que $\{N_n\}_{n\geq 0}$ qui est un processus de branchement multitype). Etant donné que le filtre particulaire simple est moins stable que son homologue convolé (cf section 7.3.2), il ne sera pas utilisé ici. L'étude complète de ce cas-test ne sera pas détaillée car les résultats qui en découlent présentent de nombreuses similitudes avec ceux du cas-test simple. Nous ne considèrerons que la situation où 5 paramètres sont à estimer (contrairement aux résultats présentés dans le tableau 7.8). A titre de comparaison, les résultats obtenus avec le filtre particulaire convolé pour le cas-test simple sont rappelés.

Globalement, les estimations des paramètres sont légèrement moins bonnes que pour le cas-test simple ce qui est normal car le système dynamique est plus complexe. En particulier, les paramètres μ et S_p sont moins bien estimés (surtout pour le filtre Rao-Blackwellisé). L'estimation des puits restent tout de même très convenable. Les temps d'exécution et le nombre d'itérations avant convergence sont bien plus élevés que ceux du cas-test simple. En effet, la suite $\{N_n\}_{n\geq 0}$ est traitée cette fois comme un processus aléatoire ce qui rajoute une grande variabilité au niveau des estimations. La convergence est donc plus difficile à obtenir. Une différence notable entre le filtre convolé et le Rao-Blackwellisé est le nombre de particules employé. Dans le cas du convolé, il faut 6000 particules pour avoir la convergence du filtre ce qui est deux fois plus que pour le cas-test

	Vraie valeur	FPC cas-test simple	FPC	FPRB
μ	9.5×10^{-2}	9.434×10^{-2}	9.048×10^{-2}	8.998×10^{-2}
S_p	2.5	2.523	2.660	2.748
P_p	0.3	0.2997	0.2971	0.2988
P_{e_1}	3	2.9984	2.8981	3.0273
P_{e_2}	2	1.9985	1.8962	2.0401
Temps d'exécution (min)	/	≈ 30	74.2	243.8
Nombre d'itérations	/	≈ 60	71	610
Nombre de particules	/	3000	6000	300

TAB. 7.8 – Résultats du cas-test de la section 7.3.6. Pour le FPC cas-test simple, les temps d'exécution et nombre d'itérations sont des valeurs moyennes. FPC = Filtre Particulaire Convolé. FPRB = Filtre Particulaire Rao-Blackwellisé.

simple. Ceci s'explique par l'augmentation de la taille de l'espace des états ($\mathcal{X}^a$ pour le cas-test simple contre $\mathcal{X}^a \times \mathcal{C}$ pour le cas-test de cette section). Il est donc normal que le temps d'exécution augmente. En ce qui concerne le filtre de Rao-Blackwell, il ne faut que 300 particules car celles-ci ne parcourent uniquement que l'espace $\mathcal{C}$ et non $\mathcal{X}^a \times \mathcal{C}$ (voir la description de la méthode dans la section 6.3.4). Paradoxalement, le temps d'exécution est bien plus important (un peu plus de trois fois supérieur). Ceci peut s'expliquer par le fait que l'état caché x_n^a des particules évolue avec un pas de filtre de Kalman sans parfum à chaque étape de la méthode. Or la correction apportée avec un pas de filtre de Kalman reste faible (surtout lors des dernières itérations de l'algorithme). Il faut donc un grand nombre d'itérations complètes avant de vérifier le critère de convergence. Le filtre de Rao-Blackwell combine ainsi les défauts des méthodes particulaires (chaque étape de l'algorithme doit être répétée autant de fois qu'il y a de particules) et ceux du filtre de Kalman (la correction apportée lors d'une étape est faible ; un grand nombre d'itérations est nécessaire pour avoir la convergence). Le filtre particulaire convolé est donc beaucoup moins gourmand en temps de calcul.

Les deux méthodes permettent de bien retrouver les états cachés du système (en particulier la vraie biomasse produite à chaque cycle de développement, voir figure 7.7).

7.3.7 Bilan des méthodes

Cette section synthétise les résultats obtenus dans toute la section 7.3. En particulier des conseils sont donnés quant au choix des méthodes à employer dans une situation précise.

Bilan quand l'évolution de la structure est déterministe

Les performances des méthodes de filtrage de Kalman sans parfum, de filtrage particulaire simple et convolé sont résumées dans le tableau 7.9.

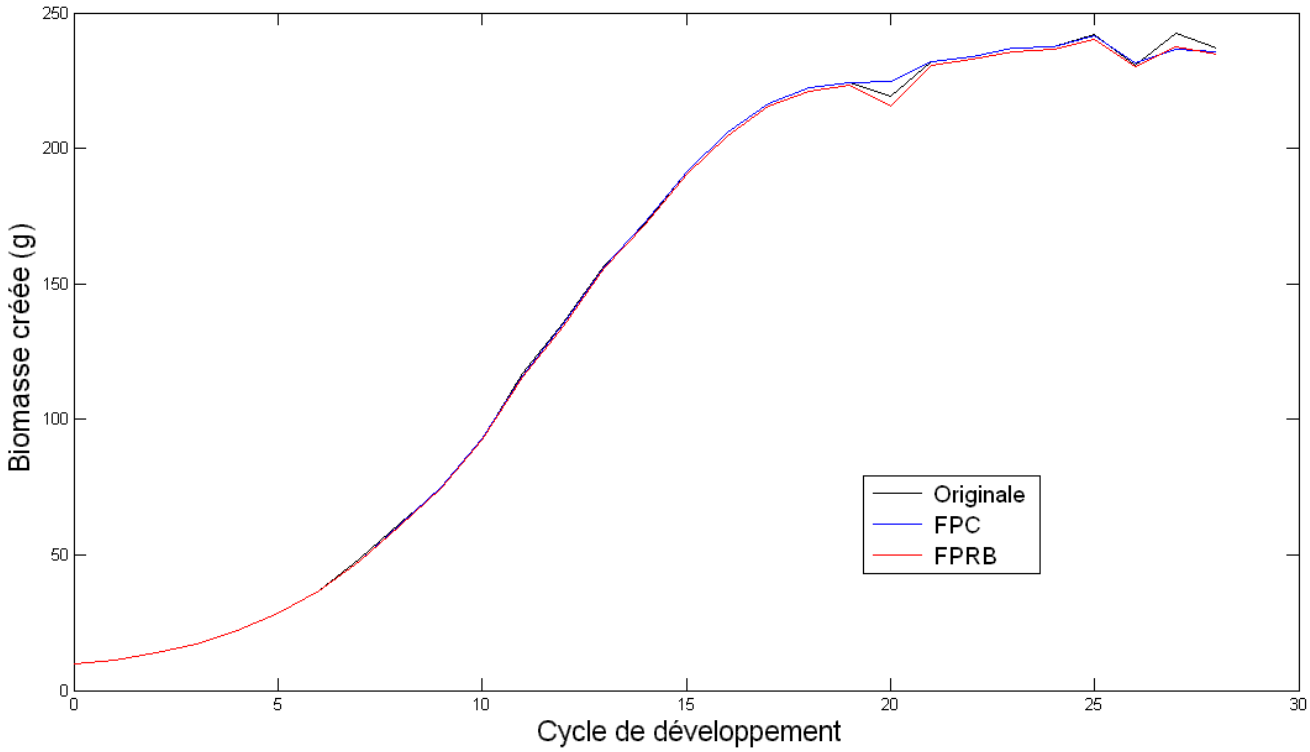

FIG. 7.7 – Estimation de la biomasse produite à chaque cycle de développement. Le filtre particulaire Rao-Blackwellisé propose une estimation un peu meilleure. FPC = Filtre Particulaire Convolé. FPRB = Filtre Particulaire de Rao-Blackwell.

	FK	FP	FPC
Estimation des paramètres	++	++	++
Estimation des états cachés	++	+	+
Convergence	++	+	+
Stabilité de l'estimation	++	.	+
Vitesse de convergence	++	+	.
Robustesse	++	+	+
Adaptativité au modèle	-	+	++
Facilité d'implémentation	.	+	++
Stabilité numérique	-	+	+

TAB. 7.9 – Bilan des méthodes pour le cas-test simple : ++ = très bon ; + = bon ; . = moyen ; - = mauvais ; - - = très mauvais. FK = Filtre de Kalman. FP = Filtre Particulaire. FPC = Filtre Particulaire Convolé.

Le filtre de Kalman propose de façon générale une meilleure estimation que ce soit au niveau des paramètres ou au niveau de la biomasse produite à chaque cycle de développement. Il est également le plus rapide à converger et propose une estimation stable à chaque exécution complète de l'algorithme puisque celui-ci est déterministe (ce qui n'est pas le cas pour les méthodes particulaires). En revanche, le filtre de Kalman trouve ses limites quand les temps d'expansion ou d'activité sont élevés ou si la structure de la plante est très ramifiée. En effet, dans ce cas, la non-linéarité du système dynamique

augmente et les équations du filtre ne sont plus valables. Les méthodes particulaires proposent des estimations qui sont tout de même bonnes et convergent plus lentement si le nombre de paramètres à estimer est élevé. Le filtre convolé est à privilégier au filtre particulaire simple car plus stable au niveau de ses estimations. Le filtre convolé peut également s'adapter à toute sorte de modèle car les hypothèses nécessaires à sa mise en place sont plus souples et il ne requiert aucun calcul complexe de densité.

En résumé, le filtre de Kalman est à privilégier pour des systèmes dynamiques faiblement non-linéaires (temps d'expansion et d'activité inférieurs ou égaux à 2 et structure de plante peu ramifiée) car il propose une très bonne estimation, converge bien et rapidement. Si le système dynamique est plus complexe, il convient alors d'utiliser le filtre particulaire convolé qui possède une grande capacité d'adaptation. Il est également stable avec de bonnes estimations.

N.B. 7.16 Le filtre de Kalman sans parfum est plus difficile à implémenter que les méthodes particulaires. En effet, les formules du filtre sont assez fastidieuses à écrire et peuvent être sources d'erreurs. En outre, la procédure de transformation sans parfum fait appel à une décomposition de Choleski qui peut s'avérer numériquement instable si les matrices en jeu sont mal conditionnées.

Bilan quand l'évolution de la structure est stochastique

Les performances des méthodes de filtrage particulaire convolé et Rao-Blackwellisé sont résumées dans le tableau 7.10.

	FPC	FPRB
Estimation des paramètres	+	+
Estimation des états cachés	.	.
Convergence	.	-
Stabilité de l'estimation	.	-
Vitesse de convergence	-	- -
Robustesse	+	+
Adaptativité au modèle	++	.
Facilité d'implémentation	++	.
Stabilité numérique	++	.

Tab. 7.10 – Bilan des méthodes pour le cas-test de la section 7.3.6 : ++ = très bon ; + = bon ; . = moyen ; - = mauvais ; - - = très mauvais. FPC = Filtre Particulaire Convolé. FPRB = Filtre Particulaire de Rao-Blackwell.

Le filtre particulaire convolé propose une estimation des paramètres un peu meilleure. Par contre, il converge beaucoup plus rapidement que le filtre Rao-Blackwellisé. Il est de plus facilement adaptable à toute sorte de modèles qu'ils soient complexes ou non. Il est donc à privilégier.

Quelques remarques

• **Filtrage et lissage** : la plante étant mesurée à un cycle T_{obs} donné, nous diposons de toutes les données $y_{0:T_{obs}}$ dès le départ. Dans cette situation, les méthodes de lissage sont à privilégier pour l'estimation des paramètres du modèle. Cependant, nous avons choisi de travailler avec des méthodes de filtrage. En effet, une itération complète d'une méthode de lissage dure approximativement deux fois plus longtemps que celle d'une méthode de filtrage du même type (par exemple filtre de Kalman et lisseur de Kalman). L'estimation fournie par la méthode de lissage est certes meilleure que celle fournie par une méthode de filtrage après une itération mais ce n'est plus le cas si l'on compare une itération de lissage avec deux itérations successives de filtrage. Ainsi le ratio qualité de l'estimation sur temps d'exécution est plus en faveur des méthodes de filtrage que de lissage dans cette thèse.

N.B. 7.17 Le lissage est à privilégier en général car il utilise toute l'information $y_{0:T_{obs}}$ à chaque étape de l'algorithme. Cependant, dans cette thèse, les algorithmes sont répétés de nombreuses fois à la suite, chacune des nouvelles itérations reprenant le vecteur de paramètres résultant de l'itération précédente. En faisant ainsi, les méthodes de filtrage utilisent également toute l'information $y_{0:T_{obs}}$ mais en décalé d'une itération à l'autre.

• **Calibrage des « autres » paramètres du modèle** : Certains paramètres fixés intervenant dans les algorithmes (par exemple, le nombre de particules M nécessaire pour permettre à l'estimateur de converger) peuvent varier suivant les modèles étudiés. Lorsque l'on travaille avec des données simulées selon un modèle de plante précis, ces paramètres sont déterminés par la pratique (on connaît alors la vraie valeur du jeu Θ_{fonc}, il suffit d'ajuster les paramètres fixés pour retrouver cette valeur par les méthodes d'estimation). Dans le cas de données obtenues à partir de plantes réelles, cette démarche ne peut plus être appliquée puisque l'on ne connaît pas a priori Θ_{fonc}. Pour contourner ce problème, il suffit d'ajuster les paramètres fixés à partir de données simulées correspondant à un modèle de plante équivalent (il faut choisir un modèle de plante avec les mêmes temps d'expansion et d'activité, les mêmes règles de production et le même nombre de paramètres dans Θ_{fonc}). Les paramètres ainsi calibrés sont bien adaptés pour traiter les vraies données.

7.4 Estimation des bruits et distribution des paramètres

Dans cette section, nous proposons tout d'abord des estimateurs pour les variances K_n et matrices de covariance R_n, $n \in \mathbb{N}$, associées respectivement aux bruits de modélisation et aux bruits de mesure (section 7.4.1). Les estimateurs sont obtenus à partir de l'estimation des états cachés augmentés du système dynamique. Il est à noter que cette procédure n'est possible que parce que l'estimation de ces états ne dépend pas du choix des bruits pourvu que ceux-ci vérifient certaines conditions (voir la section 7.3.4). Le deuxième point de cette section est la mise en place d'une procédure pour obtenir la distribution *a posteriori* du vecteur de paramètres θ_{fonc} (section 7.4.2). Cette technique repose sur l'utilisation du bootstrap paramétrique (voir Bradley and Tibshirani (1994)).

Pour toute la section, nous supposons que les bruits de modélisation $\{\omega_n,\}_{n\geq 0}$ sont des copies indépendantes d'une même variable aléatoire ω centrée et de variance K. Nous avons donc :

$$\forall n \geq 0, \qquad K_n = K_{n+1} = K.$$

7.4.1 Estimation des bruits de modélisation et de mesure

Une méthode d'estimation des variances K et σ_o^2 est proposée dans cette section dans le cas où l'évolution de la structure $\{N_n\}_{n\geq 0}$ est déterministe. A titre de rappel, K est associé au bruit de modélisation ω_n dans l'équation de production (7.7) :

$$Q_n = \Phi_n(Q_{(n-T_{act})^+}, \ldots, Q_{n-1}, N_{0\to n}, E_n, \Theta_{fonc})(1+\omega_n), \qquad n \geq 1.$$

La variance σ_o^2 est associée à l'erreur de mesure $\epsilon_n^o(i)$ de la masse d'un organe de type $o \in \mathcal{O}$ (voir équation (7.2)) :

$$M^o_{T_{obs}+1,T_{obs}+1-n}(i) = \sum_{l=0}^{\min(T_{obs}+1-n,T^o_{exp})-1} Al^{o,l}_{n+l}(Q_{n+l}, N_{0\to n+1+l}, \Theta_{fonc}) + \epsilon^o_n(i), \qquad i = 1, \ldots, \mathcal{N}^o_n.$$

Nous supposons que les états cachés du système dynamique ont été estimés par une des quatre méthodes d'inférence bayésienne à partir d'un ensemble de données $y_{0:N}$. Grâce à ces estimations, nous en déduisons :

- l'estimation $\hat{\Theta}_{fonc}$ du vecteur de paramètres Θ_{fonc} ;
- l'estimation $\hat{Q}_n$ des quantités de biomasse Q_n créées au cycle de développement $n \in \{0, \ldots, T_{obs}\}$.

L'estimateur $\hat{K}_{T_{obs}}$ de K est alors défini de la façon suivante :

Définition 7.4.1 (Estimateur de K) *L'estimateur $\hat{K}_{T_{obs}}$ de K est donné par :*

$$\hat{K}_{T_{obs}} = \frac{1}{T_{obs}-1} \sum_{n=1}^{T_{obs}} \left(\frac{\hat{Q}_n - \Phi_n(\hat{Q}_{(n-T_{act})^+}, \ldots, \hat{Q}_{n-1}, N_{0\to n}, E_n, \hat{\Theta}_{fonc})}{\Phi_n(\hat{Q}_{(n-T_{act})^+}, \ldots, \hat{Q}_{n-1}, N_{0\to n}, E_n, \hat{\Theta}_{fonc})} \right)^2$$

La terme $(\hat{Q}_n - \Phi_n(\hat{Q}_{(n-T_{act})^+}, \ldots, \hat{Q}_{n-1}, N_{0\to n}, E_n, \hat{\Theta}_{fonc}))/\hat{Q}_n$ peut s'interpréter comme une réalisation de la variable aléatoire centrée ω. L'estimateur fourni par la définition 7.4.1 est bien alors un estimateur de la variance de ω c'est-à-dire de K. De façon analogue, nous donnons un estimateur $\hat{\sigma}^o_{T_{obs}}$ pour l'écart-type σ^o associé à la mesure de la masse d'un organe de type $o \in \mathcal{O}$:

Définition 7.4.2 (Estimateur de σ^o) *L'estimateur $\hat{\sigma}^o_{T_{obs}}$ de σ^o est donné par :*

$$\left(\hat{\sigma}^o_{T_{obs}}\right)^2 = \frac{1}{T_{obs}-1} \sum_{n=1}^{T_{obs}} \mathcal{N}^o_n \left((y_n)_o - \sum_{l=0}^{\min(T_{obs}+1-n,T^o_{exp})-1} Al^{o,l}_{n+l}(\hat{Q}_{n+l}, N_{0\to n+1+l}, \hat{\Theta}_{fonc}) \right)^2.$$

avec $\mathcal{N}^o_n$ le nombre d'organes de type o créés au cycle n et $(y_n)_o = \overline{M^o_{T_{obs}+1,T_{obs}+1-n}}$ leur masse moyenne (cf section 7.1.1).

Il est alors facile d'en déduire un estimateur $\hat{R}_n$ de la matrice de covariance R_n associée à Y_n :

Définition 7.4.3 (Estimateur de R_n) *L'estimateur $\hat{R}_n$ de R_n est donné par la matrice diagonale dont les éléments diagonaux sont les composantes du vecteur $(\ldots, (\hat{\sigma}^o_{T_{obs}})^2/\mathcal{N}^o_n, \ldots)_{o\in\mathcal{O}}$.*

N.B. 7.18 L'étude théorique de ces estimateurs n'est pas faite dans cette thèse.

7.4.2 Distribution *a posteriori* des paramètres

Un fois que les matrices de covariance associées aux bruits de modélisation et de mesure sont connues (qu'elles soient fixées ou estimées), il est possible de déterminer la distribution *a posteriori* du vecteur de paramètres Θ_{fonc} en utilisant une procédure de bootstrap paramétrique. Nous pouvons alors en déduire un intervalle de confiance pour chacun des paramètres estimés. Supposons dans un premier temps que les bruits de modélisation et de mesure sont inconnus et que l'on cherche à les estimer en priorité. La procédure est alors la suivante :

1. Récolter un jeu de données $y_{0:T_{obs}}$.
2. Estimer les vecteurs d'états cachés x^a_n en choisissant $K < 10^{-2}$ et R_n une matrice diagonale dont les éléments diagonaux sont inférieurs à 10^{-2}. On obtient alors une estimation de la biomasse créée à chaque cycle de développement donnée par $\hat{Q}_n$, $n \in \{0, \ldots, T_{obs}\}$, et une estimation du vecteur de paramètres $\hat{\Theta}_{fonc}$.
3. Calculer les estimations des variances des bruits de modélisation et de mesure données par $\hat{K}$ et $\hat{R}_n$.
4. Générer un ensemble de données fictives $y^{(i)}_{0:T_{obs}}$ avec $i \in \{1, \ldots, 200\}$ à partir des bruits estimés $\hat{K}$ et $\hat{R}_n$.
5. Estimer à nouveau le vecteur de paramètres Θ_{fonc} pour chaque jeu de données $y^{(i)}_{0:T_{obs}}$, $i \in \{1, \ldots, 200\}$, en attribuant aux bruits la valeur donnée par les estimations $\hat{K}$ et $\hat{R}_n$. Le résultat est noté $\hat{\Theta}^{(i)}_{fonc}$.
6. Déterminer la distribution expérimentale de l'ensemble $\{\hat{\Theta}^{(i)}_{fonc}, i = 1, \ldots, 200\}$ et les intervalles de confiance associés à chacun des paramètres.

Rappelons une nouvelle fois que, dans les étapes 2 et 5 de la procédure décrite ci-dessus, l'estimation des états cachés augmentés dépend peu des bruits pourvu qu'ils vérifient la condition qu'on leur impose (cf section 7.3.4). En particulier, le fait que l'on impose une forme particulière à la matrice de covariance L_n associée au vecteur gaussien U_n (voir l'équation d'évolution (7.13) associée à Θ_{fonc}) n'influence en rien la méthode ci-dessus. En effet, nous n'avons besoin de L_n que pour l'estimation des paramètres et non pour générer les nouvelles données de l'étape 4 (ce sont ces nouvelles données qui vont permettre d'obtenir la distribution *a posteriori*). Dans le cas où les bruits de modélisation et de mesure sont fixés, la procédure associée est la même que ci-dessus en ne conservant que les étapes 4 à 6 et en générant les données fictives à partir des valeurs fixées des matrices de covariance.

Nous avons appliqué cette méthode au cas-test 1 pour le filtre de Kalman sans parfum, le filtre particulaire et le filtre particulaire convolé. Les estimations des bruits de modélisation et de mesure pour chacune des méthodes sont données dans le tableau 7.11.

	Vraie valeur	FK	FP	FPC
K	6.25×10^{-4}	6.453×10^{-4}	3.257×10^{-3}	3.039×10^{-3}
σ^o	2.5×10^{-2}	2.976×10^{-2}	4.781×10^{-2}	4.491×10^{-2}

TAB. 7.11 – Estimation des bruits de modélisation et de mesure pour un jeu de données du cas-test simple. FK = Filtre de Kalman. FP = Filtre Particulaire. FPC = Filtre Particulaire Convolé.

Les distributions *a posteriori* empiriques sont données par la figure 7.8 et les intervalles de confiance par le tableau 7.12.

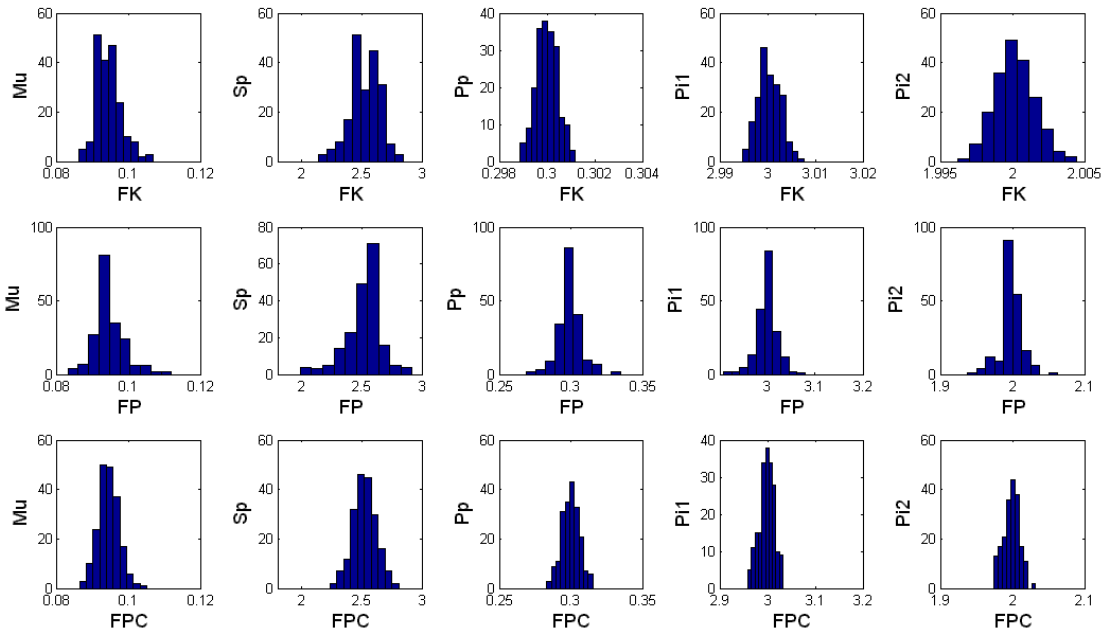

FIG. 7.8 – Histogrammes des distributions *a posteriori* du vecteur de paramètres Θ_{fonc} pour chacune des trois méthodes.

Ces résultats nous confirment à nouveau que les estimations des paramètres μ et S_p sont plus précises avec le filtre particulaire convolé. Le filtre de Kalman est plus précis pour l'estimation des puits. Notons de façon générale que la variance des estimations est toujours plus importante pour μ et S_p que pour les puits. Ces deux paramètres sont plus difficiles à estimer.

	FK	FP	FPC
μ	$[8.96 \times 10^{-2}\,;10.17 \times 10^{-2}]$	$[8.78 \times 10^{-2}\,;10.33 \times 10^{-2}]$	$[8.96 \times 10^{-2}\,;9.91 \times 10^{-2}]$
S_p	[2.28 ;2.71]	[2.22 ;2.73]	[2.35 ;2.68]
P_p	[0.299 ;0.300]	[0.284 ;0.314]	[0.289 ;0.310]
P_{e_1}	[2.99 ;3.00]	[2.96 ;3.03]	[2.97 ;3.02]
P_{e_2}	[1.99 ;2.00]	[1.97 ;2.02]	[1.98 ;2.02]

TAB. 7.12 – Intervalle de confiance à 95% pour les distributions de la figure 7.8. FK = Filtre de Kalman. FP = Filtre Particulaire. FPC = Filtre Particulaire Convolé.

N.B. 7.19 Pour les méthodes de filtrage simple et convolé, il est possible d'avoir la distribution *a posteriori* du vecteur de paramètres Θ_{fonc} en utilisant le poids des particules. Pour ce faire, il faut reprendre les trois premières étapes de la méthode précédente. Ensuite, il faut à nouveau appliquer le filtre pour l'estimation des états cachés augmentés en utilisant le même jeu de données $y_{0:N}$. Les poids finaux des particules associés à la valeur des états cachés augmentés correspondants ciblent bien la distribution *a posteriori* de Θ_{fonc}.

7.5 Application à la betterave sucrière

Nous appliquons les méthodes d'inférence bayésienne pour l'estimation du vecteur de paramètres Θ_{fonc} associé à un modèle de croissance de la betterave sucrière. Le vecteur des observations $y_{0:N}$ est construit à partir de vraies données botaniques transmises par l'ITB[1] (= Institut Technique de la Betterave). Le modèle de croissance de plante utilisé est GreenLab 1 (voir la description complète dans Lemaire et al. (2008) et Lemaire (2010)). Trois types d'organes sont pris en compte : la racine, les pétioles et les limbes. Le modèle de développement est décrit de la façon suivante :

- Au cycle 0, la germination de la graine donne une racine avec une feuille ;
- Une nouvelle feuille apparaît à chaque cycle de développement.

La figure 7.9 donne un exemple de betterave produit par le modèle GreenLab 1.

FIG. 7.9 – Betterave simulée par le modèle de croissance GreenLab 1, cf Lemaire (2010).

L'ensemble $\mathcal{O}$ associé aux organes est donné par :

$$\mathcal{O} = \{r, l, p\}.$$

Les données utilisées pour l'estimation correspondent à des valeurs moyennes sur une dizaine de plantes. Le temps d'observation vaut $T_{obs} = 51$. La racine est en expansion

[1] http ://www.itbfr.org/

permanente durant la croissance de la betterave (son temps d'expansion est choisi égal à 130). Les temps d'expansion des limbes et pétioles sont supérieurs à 1 et dépendent du rang d'insertion des feuilles sur la betterave (voir la figure 7.10). Soit alors $T_{exp}^{o,n}$, pour $o \in \{l, p\}$ et $n \in \{1, \ldots, T_{obs}\}$, le temps d'expansion d'un organe o placé au rang n et T_{exp}^{o} le maximum de ces temps d'expansion pour un organe o donné :

$$T_{exp}^{o} = \max\{T_{exp}^{o,n},\ n = 1, \ldots, T_{obs}\}.$$

Les temps d'activité des feuilles varient également en fonction de leur rang sur la plante (cf figure 7.10). Notons T_{act}^{n} le temps d'activité d'une feuille placée au rang n.

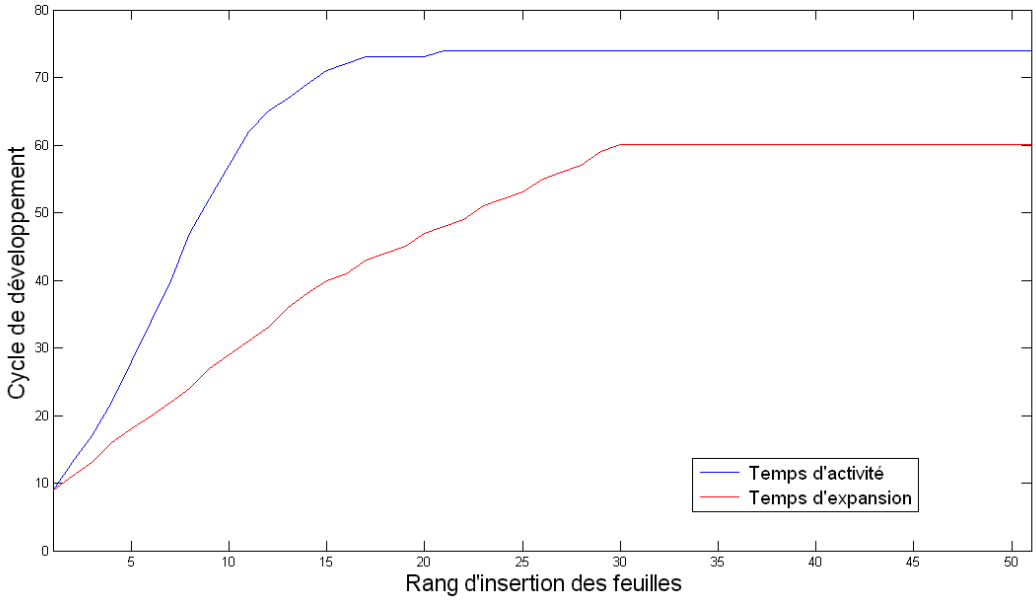

FIG. 7.10 – Temps d'expansion et d'activité des feuilles en fonction de leur rang d'insertion sur la betterave.

L'équation de photosynthèse est donnée par l'équation (2.11) du chapitre 2 :

$$Q_n = PAR_n\, RUE\, S_p \left(1 - exp\left(-\frac{k_b S_n^{tot}}{S_p}\right)\right) \qquad n \geq 1.$$

Le puits d'un organe $o \in \mathcal{O}$ placé au rang n est donné par l'équation (2.12) du chapitre 2 :

$$p_{o,n}(k) = P_o N_o \left(\frac{kT_{exp}^{o}/T_{exp}^{o,n} + 0.5}{T_{exp}^{o}}\right)^{\alpha_o - 1} \left(1 - \frac{kT_{exp}^{o}/T_{exp}^{o,n} + 0.5}{T_{exp}^{o}}\right)^{\beta_o - 1}, \quad k \in \{0, \ldots, T_{exp}^{o,n} - 1\}.$$

L'ensemble des paramètres du modèle est donc donné par le tableau 7.13. Les valeurs environnementales PAR_n sont données par la figure 7.11.

Le vecteur de paramètres Θ_{fonc} est le suivant :

$$\Theta_{fonc} = (RUE, S_p, P_p, \beta_l, \beta_p).$$

Paramètres	Description	Valeur	Nature	Unité
Par_n	Facteur environnement au cycle n	.	Mesuré	$J.m^{-2}$
K_n	Variance du bruit de modélisation	6.25×10^{-4}	Fixé	Sans Unité
σ_o	Ecart type du bruit de mesure	2.5×10^{-2}	Fixé	g
RUE	Radiation Use Efficiency	$[0.01; 0.25]$	Estimé	$g.J^{-1}$
S_p	Surface foliaire spécifique	$[0.001; 0.05]$	Estimé	m^2
k_b	Coefficient d'extinction de la loi de Beer-Lambert	0.7	Fixé	Sans Unité
P_l	Force de puits d'un limbe	1	Fixé	Sans Unité
α_l	Coefficient de la loi bêta	2.81	Fixé	Sans Unité
β_l	Coefficient de la loi bêta	$[1; 20]$	Estimé	Sans Unité
P_p	Force de puits d'un pétiole	$[0.5; 10]$	Estimé	Sans Unité
α_p	Coefficient de la loi bêta	3.64	Fixé	Sans Unité
β_p	Coefficient de la loi bêta	$[1; 20]$	Estimé	Sans Unité
P_r	Force de puits de la racine	600	Fixé	Sans Unité
α_r	Coefficient de la loi bêta	6.16	Fixé	Sans Unité
β_r	Coefficient de la loi bêta	5.57	Fixé	Sans Unité
e	Masse surfacique d'une feuille	83	Mesuré	$g.m^{-2}$
Q_0	Biomasse contenue dans la graine	0.003	Mesuré	g

TAB. 7.13 – Valeurs des paramètres pour le modèle de betterave. Le paramètre S_p rend compte de l'effet compétition. « Mesuré » signifie que le paramètre est mesuré directement à partir de données botaniques. « Fixé » signifie que la valeur des paramètres a été fixée. Enfin, « Estimé » signifie qu'il s'agit d'un paramètre à estimer à partir des données. Dans ce cas, l'intervalle indiqué dans la case valeur correspond à l'ordre de grandeur du paramètre (qui a servi à l'initialisation de l'algorithme).

Ce vecteur peut être estimé *a priori* par le filtre de Kalman sans parfum, le filtre particulaire simple et convolé car l'évolution de la structure $\{N_n\}_{n\geq 0}$ est déterministe. Etant donné que les temps d'expansion et d'activité sont très supérieurs à 1, la non-linéarité du système devient plus importante, en particulier lors des premiers stades de croissance. Il n'est donc plus possible d'appliquer le filtre de Kalman car celui-ci diverge. Le filtre particulaire convolé sera préféré à la version simple car plus stable. C'est donc le filtre convolé qui est utilisé pour l'estimation. Les résultats sont donnés dans le tableau 7.14. Etant donné le temps d'exécution important du filtre (plus de deux jours par exécution sur un ordinateur standard), seulement quatre vecteurs de paramètres Θ_{fonc} ont pu être estimés. Nous pouvons remarquer que les estimations sont moins stables surtout pour les paramètres β des puits. Ceci est dû à un nombre de particules employé trop faible par rapport à la complexité du système. En effet, nous n'avons pu utiliser que 8000 particules (ce qui correspondait à la limite de la mémoire accordée par Matlab 7.9.0 (R2009b)) alors que 12000 auraient été plus appropriées. Afin de tester l'adéquation des paramètres avec les vraies données botaniques, nous

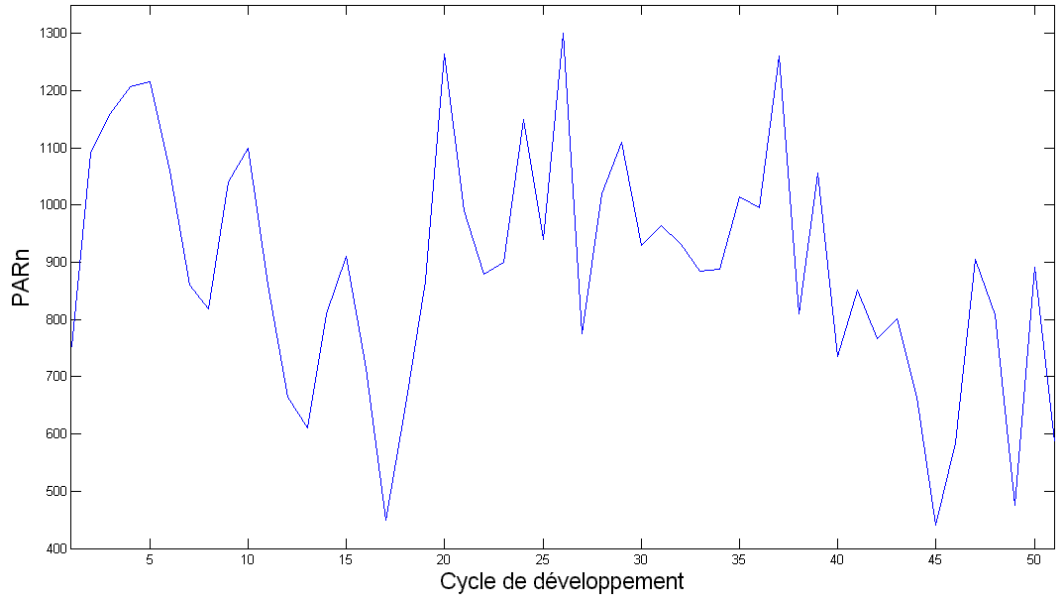

FIG. 7.11 – Valeur du PAR_n en fonction du cycle de développement.

avons simulé la croissance d'une betterave (sans bruit de modélisation ni de mesure) en utilisant la moyenne des paramètres estimés du tableau 7.14 et relevé le poids des organes ainsi obtenus. La comparaison entre valeurs simulées et valeurs réelles est donnée par la figure 7.12. Les résultats sont plutôt concluants surtout pour les limbes. Ils sont un peu moins bons pour les pétioles mais restent dans le même ordre de grandeur. Quant au poids de la racine, la valeur simulée est de 113 g contre 220 g dans la réalité. Un tel écart peut s'expliquer par l'équation de photosynthèse employée qui n'est peut être pas optimale.

N.B. : En raison du manque de temps, nous n'avons pas fait l'estimation de la distribution *a posteriori* des paramètres.

	RUE	S_p	P_p	β_l	β_p
Résultat 1	0.3714	7.231×10^{-3}	0.9814	6.6892	15.8227
Résultat 2	0.3784	8.211×10^{-3}	0.9836	8.0781	17.5643
Résultat 3	0.3698	7.988×10^{-3}	0.9823	8.1231	17.4328
Résultat 4	0.3778	7.047×10^{-3}	0.9833	7.3050	18.8205

TAB. 7.14 – Estimation des paramètres du vecteur Θ_{fonc} en exécutant 4 fois l'algorithme complet jusqu'à convergence.

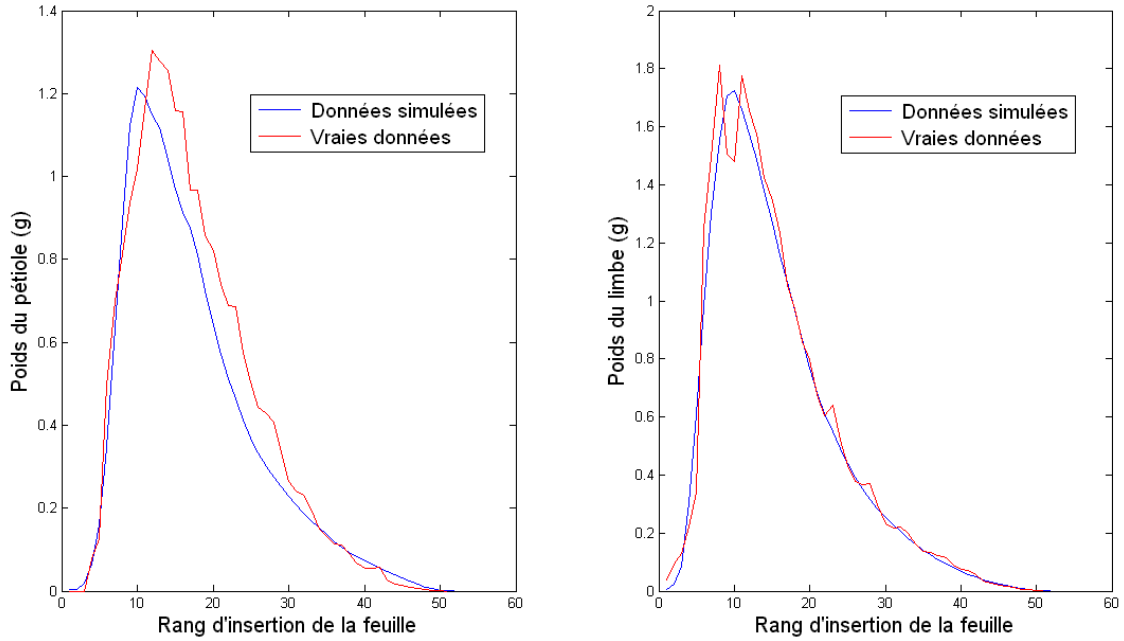

FIG. 7.12 – Comparaison des poids des pétioles et des limbes. Chaque graphique compare les poids obtenus avec les paramètres moyens estimés et les vraies données.

Chapitre 8

Conclusion et perspectives

8.1 Principaux apports

Tout au long du manuscrit, nous avons tenté de répondre aux problématiques énoncées dans l'introduction à savoir :

- l'analyse et la comparaison des différentes représentations de l'organogenèse ;
- l'étude de motifs dans la structure de la plante ;
- l'utilisation de la combinatoire concernant l'étude des structures de plante ;
- l'estimation des paramètres pour des modèles de croissance de plantes avec développement stochastique.

L'un des principaux apports de la thèse est un effort de généralisation des résultats concernant l'étude du développement stochastique de la plante et l'estimation des paramètres liés au fonctionnement. En effet, les différents travaux n'ont pas pour cible un modèle de croissance particulier mais une classe de modèles structure-fonction représentée par le métamodèle $\mathcal{M}$. Les résultats obtenus dans ce cadre sont donc valables pour tout modèle vérifiant les hypothèses énoncées dans le chapitre 2. Nous avons également présenté un nouveau modèle de développement $\mathcal{S}$ généralisant celui introduit par Kang et al. (2008). Ce dernier a été décrit sous trois angles différents : modélisation des processus (chapitre 3), processus de branchement multitypes (chapitre 3) et L-systèmes stochastiques (chapitre 5). Chacun de ces angles a ses propres avantages. Ainsi, la modélisation fait le lien entre la botanique et les mathématiques. Cette représentation de $\mathcal{S}$ est un prérequis indispensable à son étude qu'elle soit probabiliste ou combinatoire. L'approche processus de branchement nous permet de traiter des questions comme la finitude de la croissance. Enfin, la description de $\mathcal{S}$ par des L-systèmes stochastiques nous a conduit à l'élaboration d'un nouveau cadre combinatoire pour la croissance des plantes. L'une des principales nouveautés concerne l'utilisation des mots de Dyck pour coder la structure de la plante ce qui permet de conserver sa topologie. Un tel cadre couplé avec la méthode symbolique du chapitre 4 a rendu possible le calcul des distributions associées à certains motifs dans la structure de la plante (apex, structure en Y ...). L'approche combinatoire apparaît tout de même plus générale que l'approche probabiliste. En effet, les processus de branchement introduits dans le chapitre 3 ne

permettent pas l'étude de motifs puisqu'ils ne conservent pas la topologie de la plante. En revanche, nous avons montré que l'approche combinatoire était équivalente à celle des processus de branchement lorsque l'on s'intéresse aux occurrences de mots composés d'une seule lettre.

Cette thèse contribue également à l'étude probabiliste du développement de la plante. La différenciation du bourgeon apical a été modélisée par des phases-types multivariées. Ensuite, les fonctions génératrices associées au développement complet (organogenèse et différenciation) ont été établies ce qui permet par dérivation de calculer l'espérance et la variance du nombre d'organe de tout type. Ces travaux apparaissent également comme une généralisation de ceux présentés dans Kang et al. (2008). En revanche, le calcul des distributions de motifs via la méthode symbolique est entièrement nouveau. Enfin, nous avons aussi proposé une brève étude de la finitude de la croissance des plantes dont le modèle de développement est donné par $\mathcal{S}$.

Du point de vue mathématique, un autre apport majeur de la thèse est la mise en place d'une méthode symbolique pour déterminer la distribution du nombre d'occurrences d'un mot donné dans un texte généré aléatoirement par un L-système stochastique. Dans cette optique, une structure de semi-anneau reposant sur de nouveaux opérateurs union et concaténation a été établie pour l'ensemble composé des ensembles de mots pondérés construits à partir d'un alphabet donné. Nous proposons également quelques théorèmes de décomposition algébrique afin de trouver une spécification adéquate.

Du point de vue statistique, les travaux de la thèse ont contribué aux problématiques d'estimation des paramètres que ce soit pour le développement ou le fonctionnement. Concernant le développement, nous avons proposé une méthode généralisant celle de Wang et al. (2009). En effet, grâce à l'utilisation de la méthode symbolique, il est possible de faire l'estimation des paramètres non seulement à partir du nombre de phytomères mais également à partir de données botaniques liées à la topologie de la plante. Ceci permet de travailler avec des éléments de la plante qui sont plus facilement repérables (structures en Y, bouts de tige ...) et donc réduit le risque d'un comptage erroné lors de la récolte des données.

Concernant le fonctionnement, nous avons écrit $\mathcal{M}$ sous la forme d'un modèle statistique permettant l'estimation des paramètres grâce à une approche bayésienne. L'un des points forts de la méthode est d'être applicable même dans le cas où le modèle de développement est stochastique. Dans cette optique, nous avons auparavant fait une présentation dans un cadre statistique unifié de quatre méthodes d'inférence bayésiennes pour des modèles de Markov cachés. Les performances de ces méthodes ont ensuite été analysées et comparées via trois cas-tests. Finalement, un ensemble de recommandations a été donné quant à leur utilisation.

Tous ces résultats ont donné lieu à 2 publications dans des journaux avec comité de lecture, 5 publications dans les actes de conférences internationales (dont 2 dans la prestigieuse série DMTCS Proceedings) et deux rapports de recherche INRIA (voir la page 181).

8.2 Perspectives

Cette section présente quelques axes de développement possibles compte tenu des travaux réalisés jusqu'à présent.

8.2.1 Généralisation du métamodèle

Les hypothèses du métamodèle introduit dans le chapitre 2 sont assez restrictives. Afin d'élargir la classe des modèles concernés, il serait bien d'introduire la topologie comme une variable affectant le fonctionnement de la plante. En effet, dans le modèle de croissance LIGNUM par exemple (voir Perttunen et al. (1996) et Perttunen et al. (1998)), la topologie de la plante à un cycle de développement donné affecte non seulement la production de biomasse par photosynthèse mais également son allocation à l'ensemble des organes en expansion. C'est le cas également des modèles GreenLab lorsque la croissance secondaire des axes est prise en compte avec un « pipe model » (voir Shinozaki et al. (1964) et Letort (2008)). La biomasse allouée aux cernes au niveau d'un phytomère donné dépend du nombre de feuilles situées en aval du phytomère en question.

La topologie de la plante au début du cycle de développement n peut être représentée par un vecteur T_n traduisant la façon dont les organes sont reliés les uns aux autres. Deux situations sont à envisager :

- la topologie de la plante est fixée, c'est-à-dire que la suite $(T_n)_{n\geq 0}$ est connue avant même le début de la croissance : dans ce cas, le système plante peut-être décrit par l'équation de photosynthèse suivante :

  $$Q_{n+1} = \Phi_n(Q_{(n+1-T_{act})^+}, \ldots, Q_n, N_{0\to n+1}, E_{n+1}, T_{n+1}, \Theta_{fonc})(1+\omega_{n+1}), \qquad 0 \leq n,$$

  et par l'équation d'allocation :

  $$M^o_{n,k} = \sum_{l=0}^{\min(k,T^o_{exp})-1} Al^{o,l}_{n+l}(Q_{n+l}, N_{0\to n+1+l}, T_{n+l+1}, \Theta_{fonc}), \qquad 0 \leq n, \quad 0 \leq k \leq n.$$

  Etant donné que $(T_n)_{n\geq 0}$ est supposé connue à l'avance, l'étude de ce système dynamique est similaire à ceux présentés dans la thèse.

- la topologie n'est pas fixée : plusieurs causes peuvent conduire à cette situation. Nous ne considérons seulement ici que le cas où le développement est déterministe et la topologie au début du cycle $n+1$ ne dépend que de la biomasse produite au cycle n et de la topologie au début du cycle n. C'est typiquement le cas du modèle GreenLab 3(voir Mathieu et al. (2009)) avec rétroaction du fonctionnement sur l'organogenèse. Dans ce cas, le vecteur de topologie T_n vérifie une relation de récurrence du type :

  $$T_{n+1} = j_{n+1}(T_n, Q_n, \Theta_{fonc}).$$

La croissance de la plante est alors caractérisée par le système d'équations suivant :

$$\begin{cases} T_{n+1} = j_{n+1}(T_n, Q_n, \Theta_{fonc}), & 0 \leq n, \\ Q_{n+1} = \Phi_n(Q_{(n+1-T_{act})^+, \dots, Q_n, N_{0\to n+1}, E_{n+1}, T_{n+1}, \Theta_{fonc})}(1+\omega_{n+1}), & 0 \leq n, \\ M^o_{n,k} = \sum_{l=0}^{\min(k, T^o_{exp})-1} Al^{o,l}_{n+l}(Q_{n+l}, N_{0\to n+1+l}, T_{n+l+1}, \Theta_{fonc}), & 0 \leq n, \quad 0 \leq k \leq n. \end{cases}$$

Ce système peut encore être mis sous la forme d'un modèle de Markov caché en écrivant l'état caché à partir des ensembles de vecteurs $\{Q_n, n \geq 0\}$ et $\{T_n, n \geq 0\}$. Il est donc à nouveau possible d'estimer Θ_{fonc} à partir des méthodes d'inférence bayésiennes du chapitre 6.

8.2.2 Développement et exploitation de la méthode symbolique

Les travaux concernant la méthode symbolique du chapitre 4 sont encore loin d'être terminés. Il reste en effet un certain nombre de points à améliorer pour en augmenter son efficacité. Le plus important est sans aucun doute l'obtention d'une spécification appropriée pour les classes combinatoires d'intérêt. En effet, dans cette thèse, nous ne proposons pas de méthode de décomposition qui permettent d'écrire automatiquement une spécification adéquate, c'est-à-dire un ensemble d'équations algébriques qui peut se traduire directement en un système d'équations fonctionnelles faisant intervenir les fonctions génératrices d'intérêt. Deux théorèmes généraux de décomposition ont été énoncés mais ceux-ci n'aboutissent pas toujours à des concaténations non-génératrices par rapport au mot étudié (voir la section 4.4.2) et donc à des constructions admissibles. L'utilisation des règles de transformation n'est pas envisageable dans ce cas-là. Il conviendrait donc d'établir de nouveaux théorèmes décomposant les classes combinatoires avec uniquement des concaténations non-génératrices.

Ensuite, nous n'avons pas exploité tout le potentiel des relations de récurrence entre fonctions génératrices obtenues par application de la méthode. Pour le moment, nous avons uniquement établi des systèmes d'équations perrmettant de calculer de manière récursive les distributions de probabilité associées au nombre d'occurrences du mot d'intérêt ou encore les moments correspondants. Pourtant, de nombreux outils et méthodes ont été développés en analyse combinatoire pour étudier notamment le comportement asymptotique des structures combinatoires impliquées à partir du système des fonctions génératrices en question (voir Flajolet and Sedgewick (2009)).

Enfin, il faudrait multiplier les exemples illustrant la méthode symbolique au niveau des plantes en proposant de nouveaux types de motifs à étudier. De plus, il serait bien d'étendre le champs d'application de cette méthode à d'autres domaines de la biologie. Par exemple, les séquences d'ADN soumises à des mutations peuvent être vues comme des textes générés aléatoirement par des L-systèmes stochastiques. La méthode symbolique initiée dans cette thèse pourrait être envisagée pour l'étude d'occurrences de motifs dans ces séquences d'ADN en mutation.

8.2.3 Estimation des paramètres

Il reste également de nombreux travaux à accomplir concernant l'estimation des paramètres du modèle de développement stochastique $\mathcal{S}$ et du métamodèle $\mathcal{M}$. Par exemple, pour chacun des deux modèles, nous avons supposé connaître l'âge des organes composant la plante au moment où elle est observée (pour le développement, il s'agit du cas où le nombre de plantes disponibles est faible). Cette hypothèse est assez restrictive. Dans le cas d'une organogenèse stochastique avec possibilité de pause dans le développement des bourgeons, il est en général impossible de donner l'âge des organes avec exactitude ce qui peut fausser complètement les procédures d'estimation. Il faudrait donc réfléchir à de nouvelles observations ne faisant pas intervenir l'âge des organes. Concernant l'estimation des paramètres liés au développement, une solution serait de considérer des distributions de motifs dans des structures de plante qui ne sont pas caractérisées par leur âge mais par leur position dans la plante. Une telle démarche peut être envisageable avec la méthode symbolique du chapitre 4.

Concernant le modèle $\mathcal{S}$, il faudrait mettre en place une méthode découplant l'estimation des paramètres liés à la différenciation de ceux associés à l'organogenèse. En effet, les paramètres de la différenciation sont riches d'interprétation botanique. Par exemple, les temps de séjour moyens dans une classe physiologique i donnée sont représentés par les paramètres λ_i. En observant ces temps de séjour sur une population de plantes, il serait possible d'estimer λ_i, $i \in \{1, \ldots, CP_m\}$. Le fait de découpler les procédures d'estimation donnerait aux paramètres estimés un sens botanique plus fort.

Il faudrait également optimiser les algorithmes d'estimation des paramètres liés au fonctionnement. Les temps d'exécution de ces algortihmes sont trop élevés (estimation de cinq paramètres pour la betterave = 2 jours sur un ordinateur standard !). C'est en particulier le cas si la plante étudiée est très ramifiée, possède des temps d'expansion et d'activité élevés ou si le nombre de paramètres à estimer est important. En effet, les algorithmes mettent plus de temps pour se stabiliser. Dans le cas des méthodes particulaires, il se peut même que la convergence n'ait pas lieu à cause du nombre trop important de particules à employer. Plus spécifiquement, un problème se pose au niveau de l'initialisation des algorithmes (voir la section 7.3.5). Lorsque les temps d'expansion sont élevés, il est en effet nécessaire de simuler des trajectoires sur un nombre important de cycles avant de pouvoir utiliser la première observation du système. Les particules « voyagent » donc dans l'espace des états sans guidage pendant plusieurs étapes de suite ce qui entraîne une mauvaise exploration de l'espace en question. L'algorithme s'initialise donc très mal et cela peut aboutir à sa divergence. Une solution serait d'employer une approche « backward » (cf Trevezas and Cournède (2011)). Au lieu d'estimer successivement les quantités de biomasse Q_1 à $Q_{T_{obs}}$, nous faisons l'inverse, c'est-à-dire de $Q_{T_{obs}}$ à Q_1. Dans le cas « forward » (celui de la thèse), l'initialisation nécessite d'estimer les biomasses Q_1 à $Q_{T^m_{exp}}$. Il faut donc attendre T^m_{exp} étapes avant d'utiliser les premières données botaniques. Dans le cas « backward », nous pouvons utiliser les données dès l'estimation de $Q_{T_{obs}}$ (en prenant les masses des organes apparus au cycle T_{obs}). Les particules sont donc guidées dès le départ ce qui améliore les chances de convergence.

Toujours concernant le fonctionnement, l'estimation des bruits de mesure et de

modélisation est encore loin d'être satisfaisante. Les méthodes mentionnées dans la thèse reposent sur l'estimation des états cachés. Tout biais sur cette estimation en entraîne un sur celle des bruits. Ceci engendre donc des problèmes de précision quant à l'estimation de la distribution *a posteriori* du vecteur de paramètres Θ_{fonc} et en conséquence les intervalles de confiance sont erronés. Il faudrait donc mettre en place des estimateurs plus efficaces. Dans le cas des méthodes particulaires, il est possible de propager les particules selon une distribution de Wishart inverse (voir O'Hagan and Forster (2004) et Caron (2006)). Cette distribution est particulièrement adaptée pour l'estimation des matrices de covariance de lois normales. Il s'agirait donc d'une voie possible pour l'estimation des bruits. Cette méthode pourrait également être employée directement pour l'estimation de la distribution *a posteriori* des paramètres de $\mathcal{M}$ en considérant un modèle d'évolution gaussien pour le vecteur aléatoire Θ_{fonc}. Un autre axe de développement serait l'amélioration du modèle statistique associé à $\mathcal{M}$. En effet, nous avons choisi un bruit de modélisation multiplicatif et un bruit de mesure additif. Ce modèle est assez simple et pourrait être modifié pour mieux correspondre à la réalité botanique. Nous pouvons par exemple introduire du bruit dans le modèle d'allocation donné par la fonction Al (voir la section 2.2.2) ou encore améliorer le bruit de mesure en proposant des matrices de covariance traduisant des dépendances entre les mesures des différents organes (lorsque l'on coupe une feuille et que l'on sépare le pétiole du limbe, il est facile de constater que le choix de la zone de séparation influence à la fois le poids du pétiole et celui du limbe). Des techniques de sélection de modèles pourraient être envisagées pour choisir celui qui reproduit au mieux le comportement biologique de la plante.

Les méthodes bayésiennes permettent l'estimation des paramètres associés au modèle de Markov caché. Il existe cependant d'autres familles de méthodes qui ont également fait leurs preuves. Trevezas and Cournède (2011) propose une approche fréquentiste en utilisant un estimateur du maximum de vraissemblance calculé à partir d'une variante de l'algorithme *Expectation-Maximization*. Cette méthode a été appliquée avec succès sur un cas particulier du modèle de croissance GreenLab 1. Il serait intéressant de comparer les deux approches sur un même cas-test et de repérer les forces et les faiblesses de chacune d'entre elles.

Enfin, la thèse manque de mise en œuvre pratique sur données réelles. Concernant l'estimation des paramètres liés au développement, le pin (Wang et al. (2009)) peut être un beau cas-test. Il faudrait également compléter et améliorer l'étude des paramètres associés au fonctionnement de la betterave (voir le chapitre 7) avec notamment l'estimation de la distribution *a posteriori* ou encore la réduction des temps de calcul.

8.2.4 Le mot de la fin

Cette thèse apporte quelques modestes contributions au monde des modèles de croissance de plantes que ce soit en modélisation, en probabilité, en combinatoire ou en statistique. Elle s'ouvre également sur de nombreuses perspectives. Tous ces travaux montrent bien la richesse des problèmes mathématiques inhérente au domaine d'application en question. Il s'agit d'un domaine passionnant qui a encore beaucoup à offrir aux mathématiciens et qui, inversement, a beaucoup à gagner de l'utilisation des mathématiques.

Publications

• Publications dans des journaux avec comité de lecture :

C. Loi, P.-H. Cournède, and J. Françon. A Symbolic Method to Analyse Patterns in Plant Structure whose Organogenesis is Driven by a Multitype Branching Process. *Journal of Computer Science and Technology*, 2010. A paraître.

B. Pallas, C. Loi, A. Christophe, P. Cournède, and Lecoeur, J. Comparison of three approaches to model grapevine organogenesis in conditions of fluctuating temperature, solar radiation and soil water content. *Annals Of Botany*, A paraître, 2010.

• Publications dans les actes de conférences avec comité de lecture :

C. Loi and P.-H. Cournède. Generating Functions of Stochastic L-Systems and Application to Models of Plant Development. *Discrete Mathematics and Theoretical Computer Science Proceedings*, AI :325-338, 2008.

B. Pallas, C. Loi, A. Christophe, P. Cournède, and L. J. A stochastic growth model of grapevine with full interaction between environment, trophic competition and plant development. *International symposium of Plant Growth Modeling and Applications*, IEEE Computer Society (Los Alamitos, California),pages 95-102, 2009.

C. Loi, P.-H. Cournède, and J. Françon. Plants as Combinatorial Structures and Applications. In *Plant growth Modeling, simulation, visualization and their Applications (PMA09).*, IEEE Computer Society (Los Alamitos, California), pages 319-327, 2009.

C. Loi and P.-H. Cournède. A Symbolic Method to Compute the Probability Distribution of the Number of Pattern Occurrences in Random Texts Generated by Stochastic 0LSystems. *Discrete Mathematics and Theoretical Computer Science Proceedings*, AM :461-476, 2010.

C. Loi, P.-H. Cournède, and S. Trevezas. Bayesian Estimation in Functionnal-Structural Plant Models with Stochastic Organogenesis. *Proceedings of ASMDA*, 2011. Accepté.

• Rapports de recherche INRIA :

C. Loi and P.H. Cournède. Calcul des moments associés aux distributions de probabilité pour des processus de branchement. Technical report, INRIA, 2008.

C. Loi and P.H. Cournède. A Markovian framework to formalize stochastic L-systems and application to models of plant development. Technical report, INRIA, 2008.

Annexes

Annexe A

Compléments de probabilité

A.1 Fonctions génératrices

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Soit Z une variable aléatoire de $(\Omega, \mathcal{F}, \mathbb{P})$ vers l'espace mesurable $(\mathbb{N}^m, \mathcal{P}(\mathbb{N}^m))$ avec $\mathcal{P}(\mathbb{N}^m)$ l'ensemble des parties de $\mathbb{N}^m$.

Définition A.1.1 (Fonction génératrice) *La fonction génératrice de Z est une application Ψ de $[0,1]^m$ dans $[0,1]$ définie de la façon suivante :*

$$\forall (z_1, \ldots, z_m) \in [0,1]^m, \quad \Psi(z_1, \ldots, z_m) = \sum_{(k_1,\ldots,k_m)\in\mathbb{N}^m} \mathbb{P}\big(Z = (k_1, \ldots, k_m)\big) \prod_{i=1}^{m} z_i^{k_i}.$$

Comme $\sum_{(k_1,\ldots,k_m)\in\mathbb{N}^m} \mathbb{P}\big(Z = (k_1, \ldots, k_m)\big) = 1$, le rayon de convergence d'une fonction génératrice est au moins égale à 1. La fonction génératrice est donc continue sur $[0,1]^m$ (elle y est aussi convexe) et en particulier $\Psi(\mathbf{1}_m) = 1$ avec $\mathbf{1}_m$ le vecteur de taille m dont toutes les composantes sont nulles. Il s'agit également d'un outil probabiliste très utile pour le calcul des moments d'une variable aléatoire :

Théorème A.1.1 *Si la i-ème composante Z_i de Z admet un moment d'ordre $r \geq 1$ alors,*

$$\frac{\partial^r \Psi}{\partial z_i^r}(\mathbf{1}_m) = \mathbb{E}[Z_i(Z_i - 1)\ldots(Z_i - r + 1)]$$

avec $\mathbb{E}$ l'espérance d'une variable aléatoire.

Preuve Ce résultat se démontre facilement en faisant une récurrence sur r et en appliquant le lemme d'Abel (applicable puisque $\mathbb{E}[|Z_i|^r] < \infty$).

□

Corollaire A.1.2 *Si la i-ème composante Z_i de Z admet un moment d'ordre 2 alors,*

$$\mathbb{E}[Z_i] = \frac{\partial \Psi}{\partial z_i}(\mathbf{1}_m) \quad et$$
$$\mathbb{V}[Z_i] = \frac{\partial^2 \Psi}{\partial z_i^2}(\mathbf{1}_m) + \frac{\partial \Psi}{\partial z_i}(\mathbf{1}_m) - \left(\frac{\partial \Psi}{\partial z_i}(\mathbf{1}_m)\right)^2$$

avec $\mathbb{V}$ la variance d'une variable aléatoire.

Preuve Il s'agit d'une conséquence immédiate du théorème précédent en prenant les valeurs 1 et 2 pour r et en combinant les équations obtenues.

□

A.2 Phases-types multivariées

Les phase-types ont tout d'abord été développées dans les années 70 par Neuts (Neuts (1975)) avec de nombreuses applications en théorie de la fiabilité (Neuts (1981) et Assaf and Levikson (1982)). Par la suite, elles sont devenues un centre d'intérêt probabiliste qui a connu de nombreux développements mathématiques (Assaf et al. (1984), Kulkarni (1989) et plus récemment Goff (2005)).

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $(X(t))_{t\geq 0}$ un processus de Markov continue à droite sur un espace d'état fini $\mathcal{E} = \{1, \ldots, P, \Delta\}$. Supposons que $1, \ldots, P$ sont transitoires et que Δ est absorbant. Supposons également que le générateur infinitésimal Q de taille $P+1 \times P+1$ est de la forme suivante :

$$Q = \left(\begin{array}{c|c} A & -A\mathbf{1}_P \\ \hline 0 & 0 \end{array}\right)$$

avec A la sous-matrice de taille $P \times P$ du coin supérieur gauche et $\mathbf{1}_P$ un vecteur de taille P dont toutes les composantes valent 1. Soient $\Gamma_1, \ldots, \Gamma_n$ n sous-ensembles de $\mathcal{E}$ stochastiquement fermés pour X :

Définition A.2.1 *Un ensemble Γ est dit stochastiquement fermé pour le processus $(X(t))_{t\geq 0}$ si, quand X entre dans Γ, X ne quitte plus Γ :*

$$\exists t_0 \in \mathbb{R}^+, \quad X(t_0) \in \Gamma \Rightarrow \quad \forall t \geq t_o, \quad X(t) \in \Gamma.$$

Soit a la distribution initiale de X sur $\mathcal{E}$ et supposons $a(\Delta) = 0$. Soit α le vecteur de taille P défini par $\alpha = (a(1), \ldots, a(P))$. Nous définissons T_k le temps d'atteinte de Γ_k par X :

$$T_k = \inf\{j, X(j) \in \Gamma_k\}.$$

Définition A.2.2 (Vecteur aléatoire de phase-types multivariées) *Si les hypothèses suivantes sont vérifiées :*

1. $\bigcap_{k=1}^n \Gamma_k = \Delta$;
2. *l'absorption dans Δ est certaine ;*
3. $P(T_1 > 0, \ldots, T_P > 0) = 1$;

alors $(T_1, \ldots, T_P)$ est un vecteur aléatoire de phase-types multivariées avec représentation (α, A).

L'un des intérêts des phase-types est que de nombreuses quantités peuvent s'écrire de façon explicite. Soit g_k la matrice diagonale de taille $P \times P$ dont le i-ème élément diagonal vaut 1 si $i \in \Gamma_k^c$ et 0 sinon avec Γ_k^c le complémentaire de Γ_k dans $\mathcal{E}$.

Théorème A.2.1 *Soit $(T_1, \ldots, T_P)$ un vecteur aléatoire de phase-types multivariées avec représentation (α, A). Alors, pour tout $0 \leq t_1 \leq t_2 \leq \ldots \leq t_P$:*

$$P(T_1 > t_1, \ldots, T_P > t_P) = {}^t(\alpha) e^{At_1} g_1 e^{A(t_2-t_1)} \ldots g_{P-1} e^{A(t_P-t_{P-1})} g_P \mathbf{1}_P.$$

A.3 Processus de branchement de Galton-Watson multitype

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Les processus de branchement constituent une branche des probabilités encore très étudiée à l'heure actuelle et trouvent de nombreuses applications notamment en biologie moléculaire (Athreya and Ney (2004)). L'objectif est d'étudier l'évolution suivant les générations d'une population d'individus, lesquels peuvent se reproduire suivant un ensemble de règles stochastiques appelés distribution des descendants. Dans le cadre multitype, les individus peuvent être classés en m catégories numérotées de 1 à m. Un individu de type i produit des enfants de tout type suivant la distribution de probabilité suivante : $\{p_i(j) \mid j = (j_1, \ldots, j_m) \in \mathbb{N}^m\}$. $p_i(j_1, \ldots, j_k, \ldots, j_m)$ représente la probabilité pour un individu de type i de produire j_k enfants de type k pour tout $k \in \{1, \ldots, m\}$. Supposons que les individus se reproduisent de façon indépendante les uns des autres. De même, la reproduction à une génération donnée est indépendante du comportement de la population lors des générations précédentes (propriété de Markov). Soit Z_n^i la variable aléatoire donnant le nombre d'individus de type i à la génération n et soit $\mathbf{Z}_n = (Z_n^1, \ldots, Z_n^m)$. Soient $\{\xi_n[i]^{(l)} \mid (l, n) \in \mathbb{N}^2, i \in \{1, \ldots, m\}\}$ des variables aléatoires iid ayant pour distribution $\{p_i(j)\}_{j \in \mathbb{N}^m}$. $\xi_n[i]^{(l)}$ est le vecteur aléatoire dont la k-ième composante représente le nombre d'enfants de type k engendrés par l'individu l de type i de la n-ième génération. Soit $\mathbf{0}_m$ le vecteur de taille m dont toutes les composantes sont nulles.

Définition A.3.1 (Processus de Branchement de Galton-Watson Multitype) *Si le processus stochastique $(\mathbf{Z}_n)_{n \in \mathbb{N}}$ évolue suivant la relation de récurrence suivante :*

$$\forall n \in \mathbb{N}, \quad \mathbf{Z}_{n+1} = \begin{cases} \displaystyle\sum_{i=1}^{m} \sum_{l=1}^{Z_n^i} \xi_n[i]^{(l)} & si\ \mathbf{Z}_n \neq \mathbf{0}_m, \\ \mathbf{0}_m & si\ \mathbf{Z}_n = \mathbf{0}_m, \end{cases}$$

alors $(\mathbf{Z}_n)_{n \in \mathbb{N}}$ est un processus de branchement de Galton-Watson multitype.

Soit $\Psi_n[i]$ la fonction génératrice (voir l'annexe A.1) associée au vecteur aléatoire Z_n pour une lignée débutant par un individu de type i :

$$\forall (z_1, \ldots, z_m) \in [0,1]^m, \quad \Psi_n[i](z_1, \ldots, z_m) = \sum_{(k_1, \ldots, k_m) \in \mathbb{N}^m} \mathbb{P}(Z_n = (k_1, \ldots, k_m) | Z_0 = \mathbf{e}_i) \prod_{i=1}^{m} z_i^{k_i}$$

avec $\mathbf{e}_i$ le vecteur de taille m dont toutes les composantes sont nulles sauf la i-ème qui vaut 1. Soit $\mathbf{\Psi}_n = (\Psi_n[1], \ldots, \Psi_n[m])$. Alors :

Théorème A.3.1 *La fonction génératrice $\mathbf{\Psi}_n$ vérifie la relation de récurrence suivante :*

$$\forall n \in \mathbb{N}, \quad \mathbf{\Psi}_{n+1} = \mathbf{\Psi}_1 \circ \mathbf{\Psi}_n = \mathbf{\Psi}_n \circ \mathbf{\Psi}_1.$$

Preuve Pour faciliter l'écriture de la démonstration, celle-ci sera faite pour $m = 1$ (la variable aléatoire donnant le nombre d'enfants engendrés par un individu sera notée ξ). L'égalité $\mathbf{\Psi}_{n+1} = \mathbf{\Psi}_n \circ \mathbf{\Psi}_1$ est un résultat classique des fonctions génératrices puisque Z_{n+1} apparaît comme la somme de Z_n copies indépendantes de la variable aléatoire ξ. Pour l'autre égalité, nous avons pour $z \in [0, 1]$:

$$\begin{aligned}\Psi_{n+1}(z) &= \sum_{k\in\mathbb{N}} \mathbb{P}(Z_{n+1} = k) z^k \\ &= \sum_{k\in\mathbb{N}} \sum_{l\in\mathbb{N}} \mathbb{P}(Z_{n+1} = k | Z_1 = l)\mathbb{P}(Z_1 = l) z^k\end{aligned}$$

Comme $\sum_{k\in\mathbb{N}} \sum_{l\in\mathbb{N}} \mathbb{P}(Z_{n+1} = k|Z_1 = l)\mathbb{P}(Z_1 = l) = 1 < \infty$, nous pouvons appliquer le théorème de Tonelli et permuter les signes de sommation :

$$\Psi_{n+1}(z) = \sum_{l\in\mathbb{N}} \left[\sum_{k\in\mathbb{N}} \mathbb{P}(Z_{n+1} = k|Z_1 = l) z^k \right] \mathbb{P}(Z_1 = l)$$

Etant donné que $(Z_n)_{n\in\mathbb{N}}$ est une chaîne de Markov homogène et que les individus se reproduisent de façon indépendante, nous en déduisons :

$$\mathbb{P}(Z_{n+1} = k|Z_1 = l) = \mathbb{P}(Z_n = k|Z_0 = l) = \sum_{k_1+\ldots+k_l=k} \prod_{i=1}^{l} \mathbb{P}(Z_n = k_i|Z_0 = 1)$$

ce qui est le coefficient devant z^k de $(\Psi_n(z))^l$. Nous avons donc :

$$\Psi_{n+1}(z) = \sum_{l\in\mathbb{N}} \mathbb{P}(Z_1 = l) \left(\Psi_n(z)\right)^l = \mathbf{\Psi}_1 \circ \mathbf{\Psi}_n(z).$$

□

Soit E_{xt} l'évènement extinction $E_{xt} = \{\exists n \in \mathbb{N}, \quad \mathbf{Z}_n = \mathbf{0}_m\}$ et la probabilité d'extinction q_i pour la lignée issue d'un individu de type i :

$$q_i = \mathbb{P}(E_{xt}|\mathbf{Z}_0 = \mathbf{e}_i)$$

Soit $\mathbf{q} = (q_1, \ldots, q_m)$ le vecteur donnant toutes les probabilités d'extinction. Alors q vérifie :

Théorème A.3.2 *Le vecteur de probabilité d'extinction* $\mathbf{q}$ *vérifie la relation :*

$$\mathbf{q} = \mathbf{\Psi}_1(\mathbf{q}).$$

Preuve Afin de simplifier l'écriture, la démonstration sera faite pour $m = 1$ (l'extension au cas général ne pose pas de problème). La probabilité d'extinction est notée q et la distribution des descendants p. On a alors :

$$\begin{aligned} q &= \mathbb{P}(E_{xt}|Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(E_{xt}, Z_1 = k|Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(E_{xt}|Z_1 = k)\mathbb{P}(Z_1 = k|Z_0 = 1) \\ &= \sum_{k=0}^{\infty} p(k)q^k = \Psi_1(q). \end{aligned}$$

□

Soit $m_{i,j} = \mathbb{E}[Z_1^i|Z_0 = \mathbf{e}_i]$ et soit M la matrice de taille m^2 dont les composantes sont les $m_{i,j}$. Soit ρ la plus grande valeur propre de M. Supposons que $(\mathbf{Z}_n)_{n\in\mathbb{N}}$ est irréductible c'est-à-dire que, pour chaque (i,j), il existe $n_{i,j}$ tel que $(M^{n_{i,j}})_{i,j} > 0$. D'après le théorème de Perron-Frobenius, on en déduit que $\rho > 0$ et que, pour toute valeur propre λ de M, $|\lambda| < \rho$. On a alors :

Théorème A.3.3 *Soit $(\mathbf{Z}_n)_{n\in\mathbb{N}}$ un processus de branchement de GW multitype irréductible. Alors $q_i < 1$ pour tout $i \in \{1, \ldots, m\}$ si et seulement si $\rho > 1$.*

Tout comme dans le paragraphe, on peut distinguer trois cas suivant les valeurs prises par ρ :

- $\rho < 1$: le processus de branchement est sous-critique. Dans ce cas, il y a extinction de la population presque sûrement (*i.e.* $\mathbb{P}(E_{xt}|Z_0 = 1) = 1$).
- $\rho = 1$: le processus de branchement est critique. Dans ce cas, il y a extinction de la population presque sûrement (*i.e.* $\mathbb{P}(E_{xt}|Z_0 = 1) = 1$).
- $\rho > 1$: le processus de branchement est super-critique. Dans ce cas, il y a explosion de la population presque sûrement.

Annexe B

Algorithme de Levenberg-Marquardt

Soit $(d_1, d_2) \in \mathbb{N}^* \times \mathbb{N}^*$, $y \in \mathbb{R}^{d_1}$ et f une fonction Borélienne de $\mathbb{R}^{d_2}$ dans $\mathbb{R}^{d_1}$. L'objectif de l'algorithme de Levenberg-Macquardt (voir Moré (2006)) est de trouver le vecteur $p_m \in \mathcal{P} \subset \mathbb{R}^{d_2}$ tel que :

$$p_m = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \ ||y - f(p)||_{d_2}^2 = \underset{p \in \mathcal{P}}{\operatorname{argmin}} \ S(p)$$

avec $||.||_{d_2}$ la norme euclidienne sur $\mathbb{R}^{d_2}$. La procédure est alors la suivante :

Algorithme 6 : algorithme d'optimisation de Levenberg-Macquardt

---

- **Initialisation**
- Poser $p = p_0 \in \mathcal{P}$
- Poser $S^- = S(p_0)$
- Poser $condition = 1$
- Définir $\epsilon \in \mathbb{R}^+$
- **Itération**
- Tant que $condition = 1$:
  - Calcul du Jacobien de f en p : J
  - Calcul de la direction de descente q comme solution du système :

$$^tJJq = {}^tJ||y - f(p)||_{d_2}$$

  - Poser $p = p + q$
  - Si $|S(p) - S^-| < \epsilon$:
    - Poser $condition = 0$
  - Poser $S^- = S(p)$
- Poser $p_m = p$

Il est possible d'améliorer les performances de l'algorithme en modifiant l'étape d'initialisation. Au lieu de définir un point de départ $p_0 \in \mathcal{P}$, nous pouvons affiner

la zone de recherche du minimum p_m en parcourant « rapidement »l'espace des paramètres $\mathcal{P}$. Pour ce faire, nous pouvons par exemple définir un ensemble de points $P \subset \mathcal{P}$ répartit uniformément dans l'espace des paramètres et regarder pour chacun de ces points lequel donne la plus faible valeur du critère S. Ce point sera alors choisi comme point de départ de l'algorithme. Cette technique diminue grandement le risque de tomber dans un minimum local.

Annexe C

Compléments sur les méthodes d'inférence bayésienne

C.1 Equations bayésiennes pour le filtrage

On se place dans le cadre des modèles de Markov cachés. Les équations d'état sont données par le système (6.5). A l'instant n, le vecteur d'état caché est donné par X_n^a et l'observation par Y_n. Le théorème de Chapman-Kolmogorov donne l'équation de l'étape de prédiction :

$$p(x_{n+1}^a|y_{0:n}) = \int_{\mathcal{X}^a} p(x_{n+1}^a|x_n^a, y_{0:n}) p(x_n^a|y_{0:n}) \lambda(dx_n^a).$$

Dans le cadre des modèles de Markov cachés, la loi de X_{n+1}^a sachant X_n^a et $Y_{0:n}$ ne dépend que de X_n^a (voir le diagramme de dépendance 6.1). Nous en déduisons :

$$p(x_{n+1}^a|x_n^a, y_{0:n}) = p(x_{n+1}^a|x_n^a)$$

et donc :

$$p(x_{n+1}^a|y_{0:n}) = \int_{\mathcal{X}^a} p(x_{n+1}^a|x_n^a) p(x_n^a|y_{0:n}) \lambda(dx_n^a).$$

Pour l'équation de l'étape de correction, on applique d'abord la formule de Bayes à $p(x_{n+1}^a|y_{0:n+1})$ en conditionnant d'abord par Y_{n+1} puis par X_{n+1}^a :

$$\begin{aligned} p(x_{n+1}^a|y_{0:n+1}) &= p(x_{n+1}^a|y_{0:n}, y_{n+1}) \\ &= \frac{p(x_{n+1}^a, y_{n+1}|y_{0:n})}{p(y_{n+1}|y_{0:n})} \\ &= \frac{p(y_{n+1}|x_{n+1}^a, y_{0:n}) p(x_{n+1}^a|y_{0:n})}{p(y_{n+1}|y_{0:n})} \end{aligned}$$

Etant donné que Y_{n+1} dépend directement de X_{n+1}^a (voir la schéma des dépendances de la figure 6.1), on peut simplifier $p(y_{n+1}|x_{n+1}^a, y_{0:n})$ par $p(y_{n+1}|x_{n+1}^a)$. Pour le dénominateur $p(y_{n+1}|y_{0:n})$, on applique le théorème des probabilités totales en conditionnant par X_{n+1}^a. On obtient alors l'équation de l'étape de correction :

$$p(x_{n+1}^a|y_{0:n+1}) = \frac{p(x_{n+1}^a|y_{0:n}) p(y_{n+1}|x_{n+1}^a)}{\int_{\mathcal{X}^a} p(y_{n+1}|x_{n+1}^a) p(x_{n+1}^a|y_{0:n}) \lambda(dx_{n+1}^a)}$$

C.2 Transformation sans parfum (Unscented transform)

Soient X et Y deux variables aléatoires réelles reliées entre elles par une fonction f borélienne non linéaire :

$$Y = f(X).$$

L'objectif est de fournir une approximation de $E[Y] = E[f(X)]$ dans le cas où X suit une loi normale d'espérance $\overline{x}$ et de matrice de covariance Σ_x. Notons d la dimension de X. On crée d'abord un échantillon de taille $2d+1$ représentatif de la distribution de X que l'on appelle sigma-points :

$$\begin{cases} \chi^0 & = \overline{x} \\ \chi^i & = \overline{x} + \left(\sqrt{(d+\kappa)\Sigma_x}\right)_i & i \in \{1,\dots,d\} \\ \chi^{d+i} & = \overline{x} - \left(\sqrt{(d+\kappa)\Sigma_x}\right)_i & i \in \{1,\dots,d\} \end{cases}$$

avec $\kappa > -d$ et $\left(\sqrt{(d+\kappa)\Sigma_x}\right)_i$ la $i$ème ligne ou colonne de la racine carré de la matrice $(d+\kappa)\Sigma_x$. On associe à chaque point χ^i un poids ω_i de la façon suivante :

$$\begin{cases} \omega_0 & = \kappa/(d+\kappa) \\ \omega_i & = 1/2(d+\kappa) & i \in \{1,\dots,d\} \\ \omega_{i+d} & = 1/2(d+\kappa) & i \in \{1,\dots,d\} \end{cases}$$

On fait alors l'approximation suivante :

$$E[Y] = E[f(X)] \approx \sum_{i=0}^{2d} \omega_i f(\chi^{(i)}).$$

N.B. C.1 Il existe d'autres choix de sigma-points permettant des approximations d'ordre supérieur.

N.B. C.2 Si f est linéaire, alors l'approximation de $E[Y]$ est en fait exacte.

N.B. C.3 Un bon choix de κ est $d-3$ (voir Julier and Uhlmann (1997)). Dans ce cas, l'approximation du calcul de l'espérance est correcte jusqu'à l'ordre 3 (c'est-à-dire que l'espérance de Y est exacte si f est tronquée à l'ordre 3 dans son développement de Taylor).

N.B. C.4 Plutôt que de calculer la racine carré de la matrice $(d+\kappa)\Sigma_x$, on peut utiliser une décomposition de Choleski ce qui est plus rapide et plus stable numériquement.

C.3 Formule de Bayes pour le calcul séquentiel des poids des trajectoires

On se place dans le cadre des modèles de Markov cachés. Les équations d'état sont données par le système (6.5). A l'instant n, le vecteur d'état caché est donné par X_n^a et l'observation par Y_n. Le but est d'établir une équation récurrente entre $p(x_{0:n}^a|y_{0:n})$ et $p(x_{0:n+1}^a|y_{0:n+1})$. Celle-ci s'obtient en appliquant trois fois la formule de Bayes sur $p(x_{0:n+1}^a|y_{0:n+1})$ en conditionnant d'abord par Y_{n+1} ensuite par $X_{0:n}^a$ et enfin par X_{n+1}^a :

$$\begin{aligned}
p(x_{0:n+1}^a|y_{0:n+1}) &= p(x_{0:n+1}^a|y_{0:n}, y_{n+1}) \\
&= \frac{p(x_{0:n+1}^a, y_{n+1}|y_{0:n})}{p(y_{n+1}|y_{0:n})} \\
&= \frac{p(x_{0:n}^a, x_{n+1}^a, y_{n+1}|y_{0:n})}{p(y_{n+1}|y_{0:n})} \\
&= \frac{p(x_{n+1}^a, y_{n+1}|x_{0:n}^a, y_{0:n})p(x_{0:n}^a|y_{0:n})}{p(y_{n+1}|y_{0:n})} \\
&= p(x_{0:n}^a|y_{0:n})\frac{p(y_{n+1}|x_{0:n}^a, x_{n+1}^a, y_{0:n})p(x_{n+1}^a|x_{0:n}^a, y_{0:n})}{p(y_{n+1}|y_{0:n})}.
\end{aligned}$$

Etant donné que la loi de Y_{n+1} conditionnellement au passé dépend directement de X_{n+1}^a (voir la schéma des dépendances de la figure 6.1), on peut simplifier $p(y_{n+1}|x_{0:n}^a, x_{n+1}^a, y_{0:n})$ par $p(y_{n+1}|x_{n+1}^a)$. Ensuite, comme $(X_n^a)_{n\in\mathbb{N}}$ est une chaîne de Markov, on peut simplifier $p(x_{n+1}^a|x_{0:n}^a, y_{0:n})$ par $p(x_{n+1}^a|x_n^a)$. On obtient alors l'équation récurrente bayésienne pour le calcul séquentiel des poids :

$$p(x_{0:n+1}^a|y_{0:n+1}) = p(x_{0:n}^a|y_{0:n})\frac{p(y_{n+1}|x_{n+1}^a)p(x_{n+1}^a|x_n^a)}{p(y_{n+1}|y_{0:n})}.$$

C.4 Estimateurs à noyau

Le principe des estimateurs à noyau est brièvement rappelé dans ce paragraphe (voir Parzen (1962) et Silverman (1986)).

Définition C.4.1 (Noyau) *Un noyau K est une appliquation de $\mathbb{R}^d$ dans $\mathbb{R}$ vérifiant les propriétés suivantes :*

- $K(x) \geq 0, \qquad \forall x \in \mathbb{R}^d$;
- *K est symétrique :* $\forall x \in \mathbb{R}^d, \quad K(x) = K(-x)$;
- *K est bornée ;*
- $\int_{\mathbb{R}^d} K(x)dx = 1$.

On définit alors :

$$K_h(x) \stackrel{\text{def}}{=} \frac{1}{h^d}K(\frac{x}{h}), \qquad x \in \mathbb{R}^d,$$

avec $h > 0$ la taille de la fenêtre. Pour tout $h > 0$, K_h est encore un noyau.

Définition C.4.2 (Noyau de Parzen-Rozenblatt) *Un noyau K est un noyau de Parzen-Rosenblatt si :*

$$|x|^d K(x) \to 0 \qquad \textit{quand} \quad |x| \to \infty.$$

Soient $X_1, \ldots, X_N$, N variables aléatoires i.i.d. ayant la distribution de probabilité ϕ. Alors :

Définition C.4.3 (Estimateur à noyau) *L'estimateur à noyau ϕ_N de ϕ associé au noyau K est donné par :*

$$\phi_N(x) = \frac{1}{N h_N^d} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h_N}\right), \qquad x \in \mathbb{R}^d.$$

Glossaire

Les principaux concepts botaniques sont rappelés dans ce glossaire :

activité (page 36) : Une feuille est dite active à un cycle de développement donné lorsqu'elle participe à la photosynthèse durant ce même cycle.

allocation (page 35) : Étape du cycle de développement durant laquelle la biomasse produite est distribuée à l'ensemble des organes en expansion.

architecture : L'architecture d'une plante désigne la façon dont les organes sont positionnés dans l'espace ou dans le plan. Elle intègre donc la géométrie de la plante, c'est-à-dire les angles que font les organes entre eux.

biomasse (page 33) : Matière végétale composant la plante. La biomasse est produite au cours de l'étape de photosynthèse du cycle de développement.

bourgeon (page 25) : Du point de vue modélisation, le bourgeon est considéré comme l'organe de la plante responsable de la création de nouveaux phytomères lors de l'étape d'organogenèse.

classe physiologique (page 26) : Les organes de la plante peuvent être classés en différentes catégories suivant leurs caractéristiques morphologiques ou fonctionnelles. Ces catégories sont appelées classes physiologiques.

croissance : La croissance de la plante désigne l'ensemble des processus concernant le développement de sa structure et son fonctionnement biophysique au fil des cycles de développement.

cycle de développement (page 27) : Dans la thèse, la croissance de la plante est discrétisée en temps. Le cycle de développement est le pas de temps correspondant. Les processus de développement et de photosynthèse sont synchronisés sur ce pas de temps.

développement (page 23) : Le développement de la plante désigne l'ensemble des phénomènes biologiques relatifs à la mise en place de sa structure. Le développement comprend les processus d'organogenèse et de différenciation.

différenciation (page 29) : Sous l'effet du vieillissement, il est possible que le bourgeon apical d'un axe change de classe physiologique. Ce phénomène est connu sous le nom de différenciation (ou mutation) du bourgeon apical.

entrenœud (page 25) : Portion d'axe comprise entre deux nœuds.

expansion (page 35) : Un organe est dit en expansion lors d'un cycle de développement donné si sa masse augmente, c'est-à-dire s'il reçoit de la biomasse lors de l'étape d'allocation du même cycle.

fonctionnement (page 33) : Le fonctionnement de la plante désigne l'ensemble des processus de création de biomasse par photosynthèse et de son allocation.

limbe (page 25) : Partie large et aplatie de la feuille constituée de nombreuses cellules photosynthétiques.

méristème (page 24) : Tissu végétal responsable de la création de nouveaux organes. Dans la thèse, l'action des méristèmes est modélisée par les bourgeons.

métamodèle (page 23) : Modèle générique représentant une classe de modèles de croissance de plantes.

nœud (page 25) : Point d'insertion d'une feuille sur l'axe qui la porte.

organogenèse (page 23) : Étape du cycle de développement durant laquelle les bourgeons de la plante créent de nouveaux organes.

pétiole (page 25) : Partie de la feuille raccordant le limbe à l'entrenœud.

photosynthèse (page 33) : Processus biophysique durant lequel les feuilles de la plante synthétisent de la biomasse.

phytomère (page 25) : Le phytomère est l'unité élémentaire composant la structure de la plante. Il est constitué d'un entrenœud, d'une ou plusieurs feuilles et potentiellement de bourgeons, de fruits ou de fleurs.

règles de production (page 31) : ensemble des règles donnant la ou les évolutions possibles d'un bourgeon au cours d'une étape d'organogenèse.

structure (page 29) : La structure d'une plante désigne de façon indifférenciée sa topologie ou son architecture. A l'échelle de l'organe, une structure de classe physiologique i et d'âge k est la structure formée par l'ensemble des phytomères issus d'un bourgeon de CP i après k cycles de développement.

topologie (page 25) : La topologie d'une plante désigne la façon dont les organes sont agencés les uns par rapport aux autres sans considération géométrique.

Notations

Les principales notations de la thèse sont rappelées dans cette annexe.

• Variables topologiques :

- $\mathcal{B}$: ensemble des symboles associés aux bourgeons.
- $\mathcal{O}$: ensemble des symboles décrivant les éléments fixes de la plante comme les entrenœuds, les feuilles, les fleurs ou les fruits.
- b_i, e_i, f, l, p, f_r, f_l, r : symboles désignant respectivement un bourgeon de CP i, un entrenœud de CP i, une feuille, un limbe, un pétiole, un fruit, une fleur et la racine.
- $e_{i,j}$: symbole désignant un entrenœud de CP i portant des axes latéraux de CP j.
- $b^k_{\beta\to\phi}$: symbole représentant un bourgeon de CP ϕ qui est apical pour un axe d'âge k ayant débuté par un bourgeon de CP β.
- N^o_n : nombre total d'organes de type $o \in \mathcal{B}\cup\mathcal{O}$ au début du cycle de développement n.
- $N^o_n[o']$: nombre total d'organes de type $o \in \mathcal{B} \cup \mathcal{O}$ dans une structure d'âge n ayant débutée par un organe o'.
- N_n : vecteur donnant la composition complète de la plante au début du cycle n. Les composantes de N_n sont les variables N^o_n.
- B_n : vecteur dont la composante i donne le nombre de bourgeons de CP i vivants au début du cycle de développement n, c'est-à-dire $N^{b_i}_n$.
- $N_{n_1\to n_2}$: vecteur donnant l'évolution de la composition de la plante entre les cycles de développement n_1 et n_2. En d'autres termes, $N_{n_1\to n_2}$ contient les vecteurs N_k avec $k \in \{n_1, \ldots, n_2\}$.

• Croissance de la plante :

- $\mathcal{M}$: métamodèle étudié dans la thèse.
- $\mathcal{S}$: modèle de développement stochastique associé à $\mathcal{M}$.
- CP_m : nombre total de classes physiologiques.
- T^o_{exp} : temps d'expansion d'un organe $o \in \mathcal{O}$.
- T_{act} : temps d'activité d'une feuille.
- Q_n : biomasse créée au cycle n par photosynthèse.
- E_n : vecteur environnement au cycle de développement n.

- Probabilité et combinatoire :
  - $\mathbb{P}$: mesure de probabilité sur un univers Ω.
  - $\mathbb{E}$: espérance.
  - V : alphabet.
  - W : ensemble des mots construits à partir de V.
  - W^+ : ensemble des mots non vides construits à partir de V.
  - ϵ : mot vide.
  - $c(w, u)$: fonction de comptage donnant le nombre de mots u dans le mot w.
  - W_n^L : ensemble des mots générés par le 0L-système L après n étapes de production
  - π : matrice de transition associée à un 0L-système.
  - ψ_n^L : fonction génératrice d'un 0L-système L associée à un mot ou une famille de mots après n étapes de production.

- Estimation des paramètres :
  - Voir les conventions de notation du chapitre 6.
  - Θ_{org} : vecteur de paramètres associé à l'organogenèse du modèle de développement stochastique $\mathcal{S}$.
  - Θ_{org}^{GL} : vecteur de paramètres associé à l'organogenèse du modèle de croissance GreenLab 2.
  - Θ_{dif} : vecteur de paramètres associé au processus de différenciation du modèle de développement stochastique $\mathcal{S}$.
  - Θ_{dev} : vecteur formé à partir des vecteurs de paramètres Θ_{org} et Θ_{dif}.
  - Θ_{fonc} : vecteur de paramètres associé au fonctionnement de la plante.
  - T_{obs} : âge de la plante au moment de la mesure expérimentale (destructive).
  - X_n : vecteur d'état caché.
  - X_n^a : vecteur d'état caché augmenté.
  - Y_n : vecteur d'observation du système.
  - C_n : vecteur pilotant les modèles d'évolution et de mesure du système dynamique dans le cas d'un modèle à saut de Markov.
  - T_{exp}^m : maximum des temps d'expansion.
  - T : maximum des temps d'expansion et d'activité.
  - M : nombre de particules.

Bibliographie

M. Allen, P. Prusinkiewicz, and T. Dejong. Using L-Systems for Modeling Source-Sink Interactions, Architecture and Physiology of Growing Trees, the L-PEACH Model. *New Phytologist*, 166 :869–880, 2005.

B. Anderson and J. Moore. *Optimal Filtering*. Prentice-hall, 1979.

C. Andrieu, M. Davy, and A. Doucet. Efficient Particle Filtering for Jump Markov Systems. Application to Time-Varying Autoregressions. *IEEE Transactions on Signal Processing*, 51(7) :1762–1770, 2003.

M. Aono and T. Kunii. Botanical tree image generation. In *Computer Graphics and Applications*, volume 4(5), pages 10–33. IEEE, 1984.

M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on Particle Filters for On-Line Non-Linear/Non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing*, 50(2) :174–188, 2002.

D. Assaf and B. Levikson. Closure of phase-type distributions under operations arising in reliability theory. *Annals of probability*, 10 :265–269, 1982.

D. Assaf, N. Langberg, T. Savits, and M. Shaked. Multivariate phase-type distributions. *Operations Research*, 32(3) :688–702, 1984.

K. Athreya and P. Ney. *Branching Processes*. Dover Publications, 2004.

Y. Bar-Shalom, X. Li, and T. Kirubajan. *Estimation with Applications in Tracking and Navigation*, volume XLVII. 2001.

D. Barbara and T. Imielinski. Sleepers and Workoholics : Catching in Mobile Wireless Environment. In *ACM SIGMOD Proceedings*, pages 1–15. Minneapolis, 1994.

D. Barthélémy and Y. Caraglio. Plant architecture : a dynamic, multilevel and comprehensive approach to plant form, structure and ontogeny. *Annals of Botany*, 99(3) : 375–407, 2007.

D. Barthélémy, Y. Caraglio, and E. Costes. Architecture, gradients morphogénétiques et âge physiologique chez les végétaux. In J. Bouchon, editor, *Modélisation et simulation de l'architecture des végétaux*, Sciences Update, pages 89–136. INRA, 1997.

L. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics*, 37 :1554–1563, 1966.

E. Baziw. Real-Time Seismic Signal Enhancement Utilizing a Hybrid Rao-Blackwellized Particle Filter and Hidden Markov Model Filter. *IEEE Geoscience and Remote Sensing Letters*, 2(4) :418–422, 2005.

G. Blom and D. Thorburn. How many random digits are required until given sequences are obtained ? *Journal of Applied Probability*, 19 :518–531, 1982.

V. Boeva, V. Makeev, D. Papatsenko, and M. Régnier. Short Fuzzy Tandem Repeats in Genomic Sequences, Identification, and Possible Role in Regulation of Gene Expression. *Bioinformatics*, 22(6) :676–684, 2006.

N. Bouleau. *Probabilités de l'ingégieur. Variables aléatoires et simulation.* Hermann, 2002.

E. Bradley and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall/CRC, 1994.

H. Bunke and T. Caelli. *Hidden Markov Models : Applications in Computer Vision.* World Sciecntific, 2001.

F. Campillo and V. Rossi. Convolution Particle Filter for Parameter Estimation in General State-Space Models. *IEEE Transactions in Aerospace and Electronics.*, 45 (3) :1063–1072, 2009.

O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models.* Springer Science, 2005.

F. Caron. *Inférence Bayésienne pour la Détermination et la Sélection de Modèles Stochastiques.* PhD thesis, Ecole Centrale de Lille, 2006.

R. Chen, X. Wang, and J. Liu. Adaptive Joint Detection and Decoding in Flat-Fading Channels via Mixture Kalman Filtering. *IEEE Transactions on Information Theory*, 46(6) :2074–2094, 2000.

Z. Cheng, X. Zhang, and B. Chen. Simple Reconstruction of Tree Branches from a Single Range Image. *Journal of computer science and technology*, 22(6) :846–858, Nov. 2007.

A. Christophe, V. Letort, I. Hummel, P.-H. Cournède, P. De Reffye, and J. Lecoeur. A model-based analysis of the dynamics of carbon blanace at the whole plant level in *Arabidopsis thaliana* development. *Functional Plant Biology*, 35 :1147–1162, 2008.

C. Chrysaphinou and S. Papastavridis. The Occurence of Sequence of Patterns in Repeated Dependent Experiments. *Theory of Probability and Applications*, 1(35) : 167–173, 1990.

E. Costes, P. de Reffye, J. Lichou, Y. Guédon, A. Audubert, and M. Jay. Stochastic modelling of apricot growth units and branching. *Acta Horticulturae*, 313 :89–98, 1992.

E. Costes, P. Lauri, Y. Guédon, and P. de Reffye. Modelling growth of peach tree using the renewal theory. In *5th International Symposium on Orchard and Plantation Systems*, Acta Horticulturae, pages 349, 253–258, Tel-Aviv, 1993.

E. Costes, Y. Guédon, M. Jay, A. Audubert, and J. Lichou. Modeling of apricot flowers and fruits distribution in relation to shoot organization and tree architecture. In *10th International Symposium on Apricot Culture*, Izmir (Turquie), 1994. ISHS. URL `http://www-sop.inria.fr/virtualplants/Publications/1994/CGJAL94`.

P. Cournède, A. Mathieu, F. Houllier, D. Barthélémy, and P. de Reffye. Computing competition for light in the greenlab model of plant growth : a contribution to the study of the effects of density on resource acquisition and architectural development. *Annals of Botany*, 101(8), 2008.

P. Cournède, V. Letort, A. Mathieu, M. Kang, S. Lemaire, S. Trevezas, F. Houllier, and P. de Reffye. Some Parameter Estimation Issues in Functional-Structural Plant Modelling. *Mathematical Modelling of Natural Phenomena*, 6(2) :133–159, 2011.

P.-H. Cournède, M.-Z. Kang, A. Mathieu, J.-F. Barczi, H.-P. Yan, B.-G. Hu, and P. de Reffye. Structural Factorization of Plants to Compute their Functional and Architectural Growth. *Simulation*, 82(7) :427–438, 2006.

F. d'Alché Buc and N. Brunel. Estimation of Parametric Nonlinear ODEs for Biological Networks Identification. *Learning and Inference in Computational Systems Biology*, pages 61–96, 2010. MIT Press.

M. Davy. An Introduction to Statistical Signal Processing and Spectrum Estimation. *Signal Processing Methods for Music Transcription*, 2006.

P. de Reffye. *Modélisation de l'architecture des arbres par des processus stochastiques. Simulation spatiale des modèles tropicaux sous l'effet de la pesanteur. Application au Coffea robusta*. PhD thesis, Université Paris-Sud, Centre d'Orsay, 1979.

P. de Reffye and B. Hu. Relevant Choices in Botany and Mathematics for building efficient Dynamic Plant Growth Models : Greenlab Cases. In *Plant Growth Models and Applications*, pages 87–107. Tsinghua University Press and Springer, 2003.

P. de Reffye, C. Edelin, J. Françon, M. Jaeger, and C. Puech. Plant models faithful to botanical structure and development. In *Proc. SIGGRAPH 88, Computer Graphics*, volume 22(4), pages 151–158, 1988.

P. de Reffye, M. Goursat, J. Quadrat, and B. Hu. The Dynamic Equations of the Tree Morphogenesis Greenlab Model. Technical Report 4877, INRIA, 2003.

P. De Reffye, E. Heuvelink, D. Barthélémy, and P.-H. Cournède. Biological and mathematical concepts for modelling plant growth and architecture. In *Encyclopedia of Ecology*. Jorgensen, D.E. and Fath, B., elsevier edition, 2008.

Q. Dong, G. Louarn, Y. Wang, J. Barczi, and P. de Reffye. Does the structure-fonction model greenlab deal with crop phenotypic plasticity induced by plant spacing ? a case study on tomato. *Annals of Botany*, 101(8), 2008.

Q. Dong, V. Letort, G. Yan, P. De Reffye, and Z. Zhan. Modeling branching effects on source-sink relationships of the cotton plant. pages 293–300, 2009.

A. Doucet and X. Wang. Monte-Carlo Methods for Signal Processing : a Review in the Statistical Signal Processing Context. *IEEE Signal Processing Magazine*, 122(6) : 152–170, 2005.

A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

G. Duchamp, H. Kacem, and E. Laugerotte. Algebraic elimination of epsilon-transitions. *Discrete Mathematics and Theoretical Computer Science*, 7 :51–70, 2005.

J. Durand, Y. Gédon, Y. Caraglio, and E. Costes. Analysis of the plant via tree-structured statistical models : the hidden markov tree models. *New Phytologist*, 166 : 813–825, 2005.

C. Eng, C. Asthana, B. Aigle, S. Hergalant, J. Mari, and P. Leblond. A New Data Mining Approach for the Detection of Bacterial Promoters Combining Stochastic and Combinatorial Methods. *Journal of Computational Biology*, 16 (9) :1–17, 2009.

Y. Ephraim and N. Merhav. Hidden Markov Processes. *IEEE Transactions on Information Theory*, 48 :1518–1569, 2002.

G. Evensen. Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte-Carlo Methods to Forecast Error Statistics. *Journal of Geophysical Research*, 10 :143–162, 1994.

J. Fickett. Recognition of Protein Coding Regions in DNA Sequences. *Nucleic Acids Res.*, 10 :5303–5318, 1982.

S. Fine, Y. Singer, and N. Tisby. The Hierarchical Hidden Markov Model : Analysis and Applications. *Machine Learning*, 32 :41–62, 1998.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

J. Flores-Quintanilla, R. Morales-Menedez, R. Ramirez-Mendoza, L. Garza-Castanon, and F. Cantu-Ortiz. Toward a New Fault Diagnosis System for Electric Machines Based on Dynamic Probabilistic Models. In *American Control Conference. Portland, USA.*, 2005.

C. Fournier and B. Andrieu. A 3d architectural and process-based model of maize development. *Annals of Botany*, (81) :233–250, 1998.

C. Fournier and B. Andrieu. ADEL-Maize : An L-system Based Model for the Integration of Growth Process from the Organ to the Canopy. Application to Regulation of Morphogenesis by Light Availability. *Agronomy*, 19 :313–327, 1999.

J. Françon. Sur la Modélisation Informatique de l'Architecture et du Développement des Végétaux. In *2ème Colloque International : L'Arbre. Institut de Botanique, Montpellier, France*, 1990.

I. Fudos, E. Pitoura, and W. Szpankowski. On Pattern Occurrences in a Random Text. *Information Processing Letters*, 57 :307–312, 1996.

C. Gaucherel, F. Campillo, L. Misson, J. Guiot, and J. Boreux. Parameterization of a process-based tree growth model : Comparison of optimization, mcmc and particle filtering algorithms. *Environmental Modelling and Software*, 23 :1280–1288, 2008.

M. Gelfaud. Prediction of Function in DNA Sequence Analysis. *Journal of Computational Biology*, 2 :87–117, 1995.

C. Godin and Y. Caraglio. A multiscale model of plant topological structures. *Journal of Theoretical Biology*, 191 :1–46, 1998.

C. Godin and H. Sinoquet. Fonctional-strucutural plant modelling. *New Phytologist*, 166 :705–708, 2005.

M. Goff. *Multivariate discrete phase-type distributions*. PhD thesis, Washington State University, 2005.

Y. Guédon, P. Heuret, and E. Costes. Comparison methods for branching and axillayr flowering sequences. *Journal of Theoretical Biology*, 225(3) :301–325, 2003.

Y. Guédon. Estimating hidden semimarkov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3) :604–639, 2003.

Y. Guédon. Méthodes et modèles statistiques pour l'analyse de la croissance et de la structure des plantes., 2005. URL `http://www-sop.inria.fr/virtualplants/Publications/2005/Gue05b`.

Y. Guédon. Exploring the state sequence space for hidden markov and semi-markov chains. *Computational Statistics and Data Analysis*, 51(5) :2379–2409, 2007. URL `http://www-sop.inria.fr/virtualplants/Publications/2007/Gue07`.

H. Guo, V. Letort, L. Hong, T. Fourcaud, P.-H. Cournede, Y. Lu, and P. De Reffye. Adaptation of the greenlab model for analyzing sink-source relationships in chinese pine saplings. In T. Fourcaud and X. Zhang, editors, *Plant growth Modeling, simulation, visualization and their Applications (PMA06)*. IEEE Computer Society (Los Alamitos, California), 2007.

Y. Guo, Y. Ma, Z. Zhan, B. Li, M. Dingkuhn, D. Luquet, and P. de Reffye. Parameter optimization and field validation of the functional-structural model greenlab for maize. *Annals of Botany*, 97 :217–230, 2006.

F. Hallé and R. Oldeman. *Essai sur l'architecture et la dynamique de croissance des arbres tropicaux*. Masson, Paris, 1970.

F. Hallé, R. Oldeman, and P. Tomlinson. *Tropical trees and forests, An architectural analysis*. Springer-Verlag, New-York, 1978.

J. Hanan. Modelling Coton (*Gossypium hirsutum L.* with L-Systems : A Template Model for Incorporating Physiology. In *FSMP04, Montpellier, France*, June 2004.

T. Harris. *The theory of branching processes*. Springer, Berlin, 1963.

P. Heuret, D. Barthélémy, Y. Guédon, X. Coulmier, and J. Tancre. Synchronization of growth, branching, and flowering processes in the south american tropical tree cecropia obtusa (cecropiaceae). *American Journal of Botany*, 89(7) :1180–1187, 2002. URL `http://www-sop.inria.fr/virtualplants/Publications/2002/HBGCT02`. http ://www.amjbot.org/cgi/content/full/89/7/1180.

P. Heuret, Y. Guédon, N. Guérard, and D. Barthélémy. Analysing branching pattern in plantations of young red oak trees (quercus rubra l., fagaceae). *Annals of Botany*, 91(4) :479–492, 2003. URL `http://www-sop.inria.fr/virtualplants/Publications/2003/HGGB03a`. http ://aob.oxfordjournals.org/cgi/reprint/91/4/479.

E. Heuvelink. Dry matter partitioning in a tomato plant : one common assimilate pool ? *Journal of Experimental Botany*, 46(289) :1025–1033, August 1995.

E. Heuvelink. Re-interpretation of an experiment on the role of assimilate transport-resistance in partitioning in tomato. *Annals of Botany*, 78 :467–470, 1996.

E. Heuvelink. Evaluation of a dynamic simulation model for tomato crop growth and development. *Annals of Botany*, 83 :413–422, 1999.

M. Jaeger and P. de Reffye. Basic concepts of computer simulation of plant growth. *Journal of Biosciences*, 17(3) :275–291, September 1992.

F. Jelinek. Statistical Methods for Speech Recognition. *MIT Press*, 1997.

F. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996.

M. Jordan. Graphical Models. *Statistical Science*, 19 :140–155, 2004.

S. Julier and J. Uhlmann. A New Extension of the Kalman Filter to Nonlinear Systems. *International Symposium of Aerospace/Defense Sensing, Simulation and Controls*, 1997. Orlando. Fl.

S. Julier, J. Uhlmann, and H. Durrant-Whyte. A New Method for the Non-Linear Transformation of Means and Covariances in Filters and Estimators. *IEEE Transaction on Automatic Control*, 45(3) :477–482, 2000.

R. Kalman. A New Approach to Linear Filtering and Prediction Problem. *Journal of the basic engineering*, 82 :35–45, 1960.

M. Kang and P. de Reffye. A mathematical approach estimating source and sink functioning of competing organs. In J. Vos, L. Marcelis, P. de Visser, P. Struik, and J. Evers, editors, *Functional-structural plant modelling in crop production, Wageningen*, volume Chapter 06, pages 70–80. Springer, 2007.

M. Kang, P.-H. Cournède, J.-P. Quadrat, and P. de Reffye. A stochastic language for plant topology. In T. Fourcaud and X. Zhang, editors, *Plant growth Modeling, simulation, visualization and their Applications.* IEEE Computer Society (Los Alamitos, California), 2007.

M. Kang, L. Yang, B. Zhang, P. de Reffye, D. Auclair, and B.-G. Hu. Correlation Between Dynamic Tomato Fruit-Set and Source-Sink Ratio : a Common Relationship for Different Plant Densities and Seasons ? *Annals Of Botany*, pages 1–11, 2010. doi : 10.

M.-Z. Kang, P. de Reffye, J.-F. Barczi, B.-G. Hu, and F. Houllier. Stochastic 3d tree simulation using substructure instancing. In B. Hu and M. Jaeger, editors, *Plant Growth Models and Applications PMA03*, pages 154–168. Tsinghua University Press and Springer (Beijing, China), 2003.

M.-Z. Kang, P.-H. Cournède, J. Le Roux, P. de Reffye, and B.-G. Hu. Theoretical study and numerical simulation of a stochastic model for plant growth. In *CARI04, Tunisia*, 2004.

M.-Z. Kang, P.-H. Cournède, P. de Reffye, D. Auclair, and B.-G. Hu. Analytical study of a stochastic plant growth model : application to the greenlab model. *Mathematics and Computers in Simulation*, 78(1) :57–75, 2008.

S. Karlin and C. Macken. Some Statistical Problems in the Assessment of Inhomogeneities of DNA Sequence Data. *Journal of American Statistical Association*, 86 :27–35, 1991.

C. Kim and C. Nelson. State-Space Models with Regime Switching : Classical and Gibbs-Sampling Approaches with Applications. *MIT Press*, 1999.

G. Kitagawa. Monte-Carlo Filter and Smoother for Non-Gaussian Non-Linear State-Space Models. *Journal of Computational and Graphical Statistics*, 5(1) :1–25, 1996.

O. Klima and L. Polak. On varieties of meet automata. *Theoretical Computer Science*, 407 :278–289, 2008.

D. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley Professional, 1997 (3rd edition).

A. Kong, J. Liu, and W. Wong. Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association*, 89(425) :278–288, 1994.

T. Koski. *Hidden Markov Models for Bioinformatics*. Kluwer, 2001.

P. Kruszewski and S. Whitesides. A General Random Combinatorial Model of Botanical Trees. *Journal of Theoretical Biology*, 191 :221–236, 1998.

V. Kulkarni. A new class of multivariate phase-type distributions. *Operations research*, 37 :151–158, 1989.

V. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall, London, 1995.

W. Kurth. Specifications of morphological models with l-systems and relational growth grammars. *Journal of Interdisciplinary Image Science*, 5, 2007.

W. Kurth. Morphological models of plant growth : Possibilities and ecological relevance. *Ecological Modelling*, 75/76 :299–308, 1994a.

W. Kurth. *Growth grammar interpreter GROGRA 2.4 : A software tool for the 3-dimensional interpretation of stochastic, sensitive growth grammars in the context of plant modelling. Introduction and Reference Manual*. Berichte des Forschungszentrums Waldokosysteme der Universitat Gottingen, Ser. B, Vol. 38, 1994b.

W. Kurth. Some new formalisms for modelling the interactions between plant architecture, competition and carbon allocation. In *4th workshop on individual-based structural and functional models in ecology*. Wallenfels, Bayreuther forum ökologie, 9 1996.

W. Kurth and B. Sloboda. Growth grammars simulating trees-an extension of l-systems incorporating local variables and sensitivity. *Silva Fennica*, 31(3) :285–295, 1997.

V. Le Chevalier. *Cadres formels pour la simulation des peuplements hétérogènes de plantes en compétition pour les ressources.* PhD thesis, Ecole Centrale Paris, 2010.

X. Le Roux, A. Lacointe, A. Escobar-Gutiérrez, and S. Le Dizès. Carbon-based models of individual tree growth : A critical appraisal. *Annals of Forest Sciences*, 58 :469–506, 2001.

S. Lemaire. *Système Dynamique de la Croissance et du Développement de la Betterave Sucrière (Beta vulgaris L.)*. PhD thesis, AgroParisTech, 2010.

S. Lemaire, F. Maupas, P.-H. Cournède, and P. de Reffye. A Morphogenetic Crop Model for Sugar-Beet (*beta vulgaris l.*). In *International Symposium on Crop Modelling And Decision Support : ook ISCMDS 2008, April 19-22, 2008, Nanjing, China*, 2008.

V. Letort. *Multi-scale analysis of source-sink relationships in plant growth models for parameter identification. Case of the GreenLab model.* PhD thesis, Ecole Centrale Paris, 2008.

V. Letort, P.-H. Cournède, A. Mathieu, P. de Reffye, and T. Constant. Parameter identification of a functional-structural tree growth model and application to beech trees (*Fagus sylvatica*). *Functional Plant Biology*, 35 :951–963, 2008a.

V. Letort, P. Mahe, P.-H. Cournède, P. de Reffye, and B. Courtois. Quantitative genetics and functional-structural plant growth models : simulation of quantitative trait loci detection for models parameters and application to potential yield optimization. *Annals of Botany*, tome 101 (8), 2008b.

V. Letort, P. Heuret, P. Zalamea, E. Nicolini, P. de Reffye, and B. Courtois. Analysis of cecropia sciadophylla morphogenesis based on a sink-source dynamic model. In *3rd international symposium on Plant Growth and Applications (PMA09), Beijing, China.*, pages 10–17. IEEE Computer Society (Los Alamitos, California), 2009.

A. Lindenmayer. Mathematical models for cellular interactions in development. i. filaments with one-sided inputs. *Journal of Theoretical Biology*, 18 :280–289, 1968.

J. Liu and M. West. Combined Parameter and State Estimation in Simulation-Based Filtering. *Sequential Monte Carlo Methods in Practice.*, pages 197–223, 2001.

C. Loi and P.-H. Cournède. A Markovian framework to formalize stochastic L-systems and application to models of plant development. Technical report, INRIA, 2008.

C. Loi and P.-H. Cournède. Generating Functions of Stochastic L-Systems and Application to Models of Plant Development. *Discrete Mathematics and Theoretical Computer Science Proceedings*, AI :325–338, 2008.

C. Loi and P.-H. Cournède. A Symbolic Method to Compute the Probability Distribution of the Number of Pattern Occurrences in Random Texts Generated by Stochastic 0L-Systems. *DMTCS*, AM :461–476, 2010.

C. Loi, P.-H. Cournède, and J. Françon. Plants as Combinatorial Structures and Applications. In *Plant growth Modeling, simulation, visualization and their Applications (PMA09).*, pages 319–327. IEEE Computer Society (Los Alamitos, California), 2009.

C. Loi, P.-H. Cournède, and J. Françon. A Symbolic Method to Analyse Patterns in Plant Structure whose Organogenesis is Driven by a Multitype Branching Process. *Journal of Computer Science and Technology*, 2010. A paraître.

C. Loi, P.-H. Cournède, and S. Trevezas. Bayesian Estimation in Functionnal-Structural Plant Models with Stochastic Organogenesis. *Proceedings of ASMDA*, 2011. Accepté.

G. Lopez, R. Favreau, C. Smith, E. Costes, P. Prusinkiewicz, and T. DeJong. Integrating Simulation of Architectural Development and Source-Sink Behaviour of Peach Trees by Incorporating Markov Chains and Physiological Organ Function Submodels into L-PEACH. *Functional Plant Biology*, 35 :761–771, 2008.

T. Luczak and W. Szpankowski. A Suboptimal Lossy Data Compression Based on Approximate Pattern Matching. *IEEE transactions on Information Theory*, 43 (5) : 1439 – 1451, 1997.

X. Luo and I. Moroz. Ensemble Kalman Filter with the Unscented Transform. *Physica D*, pages 549–562, 2009.

Y. Ma, A. Wubs, Y. Guo, A. Mathieur, H. E., J. Zhu, B. HU, P. Cournède, and P. De Reffye. Simulation of Fruit Set and Trophic Competition and Optimization of Yield Advantages in Six *Capsicum*.

L. Marcelis, E. Heuvelink, and J. Goudriaan. Modelling of biomass production and yield of horticultural crops : a review. *Scientia Horticulturae*, 74 :83–111, 1998.

A. Mathieu. *Essai sur la modélisation des interactions entre la croissance d'une plante et son développement dans le modèle GreenLab*. PhD thesis, Ecole Centrale Paris, 2006.

A. Mathieu, P. Cournède, V. Letort, D. Barthélémy, and P. de Reffye. A dynamic model of plant growth with interactions between development and functionnal mechanisms to study plant structural plasticity related to trophic competition. *Annals of Botany*, 2009.

C. Mode. *Multitype branching processes : Theory and applications*. American Elsevier Publishing Co. Inc, New York, 1971.

J. Monteith. Climate and the efficiency of crop production in britain. *Philosophical Transactions of the Royal Society of London*, Series B Biological Sciences 281 :277–294, 1977.

J. Moré. *The Levenberg-Marquardt Algorithm : Implementation and Theory*, pages 105–116. Springer, springer berlin / heidelberg edition, 2006. ISBN : 978-3-540-08538-6.

M. F. Neuts. Probability distributions of phase type. *Liber Amicorum Prof. Emeritus H. Florin*, pages 176–206. University of Louvain, Belgium, 1975.

M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. John Hopkins University Press, Baltimore, 1981.

E. Nicolini. *Approche morphologique du développement du hêtre (Fagus sylvatica L.)*. PhD thesis, University of Montpellier II, France, 1997.

A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics : Bayesian Inference*. Arnold, 2nd édition, 2004. ISBN 0-340-80752-0.

B. Pallas, C. Loi, A. Christophe, P. Cournède, and L. J. A stochastic growth model of grapevine with full interaction between environment, trophic competition and plant development. *International symposium of Plant Growth Modeling and Applications*, pages 95–102, 2009.

B. Pallas, C. Loi, A. Christophe, P. Cournède, and L. J. Comparison of three approaches to model grapevine organogenesis in conditions of fluctuating temperature, solar radiation and soil water content. *Annals Of Botany*, In press, 2010.

E. Pardoux. *Processus de Markov et Applications*. Dunod, 2007.

E. Parzen. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

J. Perttunen, R. Sievänen, E. Nikinmaa, H. Salminen, H. Saarenmaa, and J. Väkevä. Lignum : a tree model based on simple structural units. *Annals of Botany*, 77 :87–98, 1996.

J. Perttunen, R. Sievänen, and E. Nikinmaa. Lignum : a model combining the structure and the functioning of trees. *Ecological Modelling*, 108 :189–198, 1998.

D. Pham. Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems. *Monthly Weather Review*, 129(5) :217–244, 2001.

M. Pitt and N. Shepard. Auxiliary Variable Based Particle Filters. *Sequential Monte Carlo Methods in Practice.*, pages 273–293, 2001.

B. Prum, F. Rodolphe, and E. Turckheim. Finding Words with Unexpected Frequencies in DNA Sequences. *Journal of Royal Statistical Society*, B(57) :205–220, 1995.

P. Prusinkiewicz. Modeling plant growth and development. *Current opinion in plant biology*, 7(1) :79–84, 2004.

P. Prusinkiewicz and A. Lindenmayer. *The Algorithmic Beauty of Plants*. Springer-Verlag, New-York, 1990.

M. Quach, N. Brunel, and F. d'Alché Buc. Estimating Parameters and Hidden Variables in Non-Linear State-Space Models Based on ODEs for Biological Networks Inference. *Bioinformatics*, 23(23) :3209–3216, 2007.

L. Quan, P. Tan, G. Zeng, L. Yuan, J. Wang, and S. Kang. Image Based Plant Modeling. *ACM Trans. Graph.*, 25(3) :599–604, 2006.

L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings IEEE*, 77 :257–285, 1989.

M. Régnier. A Unified Approach to Word Occurrence Probabilities. *Discrete Applied Mathematics*, 104 :259–280, 2000.

M. Régnier and A. Denise. Rare Events and Conditional Events on Random Strings. *DMTCS*, 6 :191–214, 2004.

M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4) :631–649, 1998. URL `http://algo.inria.fr/papers/other/ReSz97b.ps.gz`.

J. Riordan. *An Introduction to Combinatorial Analysis*. Courier Dover Publications, 2002.

B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter : Particle Filters for Tracking Applications*. Artech House, 2004.

S. Robin. *Répartition de Motifs dans les Séquences d'ADN*. 2002. Habilitation à diriger des recherches - Université d'Evry Val d'Essonne.

S. Robin and J. Daudin. Exact Distribution of Word Occurrences in a Random Sequence of Letters. *Journal of Applied Probability*, 36 :179–193, 1999.

S. Robin, J. Daudin, H. Richard, M. Sagot, and S. Schbath. Occurrence Probability of Structured Motifs in Random Sequences. *Journal of Computational Biology*, 9 (6) : 761–773, 2002.

G. Rozenberg and A. Salomaa. *The Mathematical Theory of L-systems*. Academic Press, New York, 1980a.

G. Rozenberg and A. Salomaa. *The Book of L*. Springer, Berlin, 1980b.

G. Rozenberg and A. Salomaa. *Lindenmayer Systems : Impacts on Theoretical Computer Science, Computer Graphics and Developmental Biology*. Springer, Berlin, 1992.

S. Sabatier and D. Barthélémy. Growth dynamics and morphology of annual shoots according to their architectural position in young cedrs atlantica (endl.) manetti ex carrière (pinaceae). *Annals of Botany*, 84 :387–392, 1999.

K. Shinozaki, K. Yoda, K. Hozumi, and T. Kira. A quantitative analysis of plant form - the pipe model theory i. basic analysis. *Japanese Journal of Ecology.*, 14 :97–105, 1964.

R. Sievänen, E. Nikinmaa, P. Nygren, H. Ozier-Lafontaine, J. Perttunen, and H. Hakula. Components of a functional-structural tree model. *Annals of Forest Sciences*, 57 :399–412, 2000.

B. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

S. Säkkä, A. Vehtari, and J. Lampinen. Rao-Blackwellized Monte-Carlo Data Association for Multiple Target Tracking. In *International Conference on Information Fusion (FUSION 2004)*, 2004. Stockholm, Suède.

A. Smith. Plants, fractals and formal languages. *Computer Graphics (SIGGRAPH 84 Conference Proceedings)*, 18(3) :1–10, 1984.

H. Sorenson. *Kalman Filtering : Theory and Applications*. IEEE Press, 1985.

W. Szpankowski. Asymptotic Properties of Data Compression and Suffix Trees. *IEEE transactions on Information Theory*, 39 :1647–1659, 1993.

W. Taylor. Small Sample Properties of a Class of Two-Stage Aitken Estimator. *Economica*, 45(2) :497–508, 1977.

S. Trevezas and P. Cournède. A Sequential Monte Carlo Approach for MLE in a Plant Growth Model. 2011. soumis.

X. Viennot, G. Eyrolles, N. Janay, and D. Arques. Combinatorial analysis of ramified patterns and computer imagery of trees. In *Computer Graphics*, volume 23(3), pages 31–40. 1989.

J. Vose, N. Sullivan, B. Clinton, and P. Bolstad. Vertical leaf area distribution, light transmittance, and application of the beer-lambert law in four mature hardwood stands in the southern appalachians. *Canadian Journal of Forest Research*, 25 :1336–1343, 1995.

D. Wallach, D. Makowski, and J. Jones. *Working with Dynamic Crop Models - Evaluation, Analysis, Parameterization and Applications.* Elsevier B.V., 2006.

E. Wan and R. Van Der Merwe. The Unscented Kalman Filter for Non-Linear Estimation. *IEEE Symposium 2000, lake Louise, Alberta, Canada*, 2000.

F. Wang, M. Kang, Q. Lu, H. Han, V. Letort, Y. Guo, R. De Reffye, and B. Li. Calibration of the Topological Development in the Procedure of Parametric Identification : Application of the Stochastic Greenlab Model for *Pinus Sylvestris* var. *Mongolica*. In *3rd international symposium on Plant Growth and Applications (PMA09), Beijing, China.*, pages 26–33. IEEE Computer Society (Los Alamitos, California), 2009.

F. Wang, M. Kang, Q. Lu, V. Letort, H. Han, Y. Guo, P. De Reffye, and B. Li. A Stochastic Model of Tree Architecture and Biomass Partioning : Application to Mongolian Scots Pines. *Annals Of Botany*, pages 1–12, 2010. doi : 10.1093/aob/mcq218. URL `http://hal.inria.fr/hal-00546767/en`.

J. Warren-Wilson. *Control of Crop processes*, pages 7–30. Rees, AR and Cockshull, KE and Hand, DW and Hurd, RG, london : academic press edition, 1972.

J. White. The plant as a metapopulation. *Annu. Rev. Ecol. Syst*, 10 :109–145, 1979.

Y. Wu, G. Hua, and T. Yu. Switching observation models for contour tracking in clutter. In *Dans Aerospace Conference*, 2003.

H. Xu, N. Gosset, and B. Chen. Knowledge Based Modeling of Laser Scanned Trees. *In SIGGRAPH' 05 : ACM SIGGRAPH 2005 Sketches*, page 124, 2005.

H. Yan, M. Kang, P. De Reffye, and M. Dingkuhn. A dynamic, architectural plant model simulating resource-dependent growth. *Annals of Botany*, 93 :591–602, 2004.

X. Yin, P. Stam, M. Kropff, and A. Schapendonk. Crop modeling, qtl mapping, and their complementary role in plant breeding. *Agronomy Journal*, 95 :90–98, 2003.

X. Zhao, P. de Reffye, F. Xiong, B. Hu, and Z. Zhan. Dual-scale automaton model of virtual plant growth. *Chinese Journal of Computers*, 24(6) :608–615, 2001.

X. Zhao, P. de Reffye, D. Barthélémy, and B. Hu. Interactive simulation of plant architecture based on a dual-scale automaton model. In *Plant Growth Models and Applications*, pages 144–153. Tsinghua University Press and Springer, 2003.

C. Zhu, X. Zhang, M. Jaeger, and Y. Wang. Cluster-Based Construction of Tree Crown from Scanned Data. In *Plant growth Modeling, simulation, visualization and their Applications (PMA09).*, pages 352–359. IEEE Computer Society (Los Alamitos, California), 2009.