HSMM piloté par les observations pour l'estimation de la dynamique des adventices

Hanna Bacave[1] & Nikolaos Limnios[2] & Nathalie Peyrard[1]

[1] *INRAE, UR MIAT, Université de Toulouse, Castanet-Tolosan, France.* {*hanna.bacave, nathalie.peyrard*}*@inrae.fr*

[2] *Sorbonne University Alliance, Unversité de Technologie de Compiègne, LMAC, France. nikolaos.limnios@utc.fr*

Résumé. Les adventices sont des plantes qui poussent spontanément dans les parcelles agricoles et qui entrent en compétition avec les cultures. Leur dynamique repose sur la colonisation et la dormance. La banque de graines n'étant jamais observée de manière naturelle, une modélisation de cette dynamique a été proposée dans le cadre des Hidden Markov Models (HMM). Ce modèle, appelé Observation Driven-HMM (OD-HMM) étend les HMM au cas où les probabilités de transition dépendent de l'observation courante pour tenir compte des nouvelles graines produites qui entrent dans la banque de graines. Cependant, pour plus de réalisme sur la distribution de la survie de la banque de graines, le cadre naturel serait celui des Hidden Semi-Markov Models (HSMM). Néanmoins la notion de durée de séjour dans l'état caché n'est plus adaptée dès lors que l'observation influence la chaîne cachée à chaque instant. En nous appuyant sur les deux cadres OD-HMM et HSMM, nous proposons un nouveau modèle général : l'OD-HSMM, permettant à la fois de tenir compte d'une influence des données sur la chaîne cachée et de s'affranchir de la loi du temps de séjour géométrique. Nous en présentons une version paramétrique à partir des paramètres clés de la dynamique d'une espèce adventice et nous discutons différentes approches pour leur estimation.

Mots-clés. HSMM, OD-HMM, algorithme ABC, algorithme EM, adventices

Abstract. Weeds are plants that grow spontaneously in agricultural plots and compete with crops. Their dynamics are based on colonization and dormancy. As the seed bank is never observed, a model of these dynamics has been proposed within the HMM framework. This model, called OD-HMM, extends HMM to the case where transition probabilities depend on current observation, to take account of newly produced seeds entering the seed bank. However, for more realism on the survival distribution of the seed bank, the natural framework would be Hidden Semi-Markov Models (HSMM). However, the notion of sojourn time in the hidden state is no longer appropriate, since observation influences the hidden chain at every instant. Combining the OD-HMM and HSMM frameworks, we propose a new general model, the OD-HSMM, which allows us to take into account the influence of data on the hidden chain, while at the same time going beyond the geometric distribution of sojourn time. We present a parametric version based on key parameters of the dynamics of a weed species, and discuss different approaches for their estimation.

Keywords. HSMM, OD-HMM, ABC algorithm, EM algorithm, weeds

1 Introduction

Les adventices sont des plantes qui poussent spontanément dans les parcelles agricoles et qui entrent en compétition avec les cultures, entraînant parfois une baisse du rendement. Dans le même temps, elles constituent un refuge pour les insectes pollinisateurs, comme les abeilles, ce qui leur confèrent un rôle important dans l'équilibre écologique. Ainsi, il est important de comprendre et donc de modéliser leur développement pour trouver un mode de gestion offrant un compromis entre conservation et éradication des plantes. La difficulté vient du fait qu'un élément important de la dynamique des plantes en général et des adventices en particulier est invisible : la banque de graines dans le sol. En général on ne dispose de données que sur la flore levée. Dans ce contexte avec données manquantes, ou cachées, une modélisation de type modèle de Markov caché (Hidden Markov Model, HMM) a naturellement été proposée (Pluntz et al., 2018). Ce modèle étend les HMM au cas où les probabilités de transition dépendent de l'observation courante pour tenir compte des nouvelles graines produites qui entrent dans la banque de graines. Pour cette raison il est appelé Observation-Driven HMM (OD-HMM, Bacave et al., 2023).

Cependant, dans l'OD-HMM la chaîne cachée vérifie toujours l'hypothèse markovienne et son utilisation implique donc de supposer que la durée de vie de la banque de graines est distribuée selon une loi géométrique, réduisant considérablement la probabilité d'avoir une durée de vie de la graine importante. Or, dans la réalité les graines peuvent survivre plusieurs dizaines d'années dans le sol (Baskin and Baskin, 2014). Le cadre naturel pour une loi du temps de séjour plus générale est le cadre des Hidden Semi Markov Models (HSMM, Barbu and Limnios, 2008; Yu, 2016). En nous appuyant sur ce cadre, nous proposons donc un nouveau modèle général, l'OD-HSMM, permettant à la fois de tenir compte d'une influence des données sur la chaîne cachée et de sortir de la loi du temps de séjour géométrique (Section 2).

Nous présentons ensuite une version paramétrique de l'OD-HSMM permettant d'intégrer les paramètres en jeu dans la dynamique d'une espèce adventice, à savoir ses probabilités de colonisation, germination, grenaison ainsi que de survie (Section 3).

Enfin, nous explorons deux approches pour l'estimation des paramètres. La première repose sur l'algorithme EM (Dempster et al., 1977), classiquement utilisé pour les modèles à variables cachées. Nous l'appliquons ici à l'OD-HSMM non paramétrique car l'étape M est alors explicite pour la mise à jour des probabilités de transition et d'émission puis nous rajoutons une étape de minimisation pour identifier les paramètres adventices à partir de ces distributions. La seconde repose sur l'algorithme Approximate Bayesian Computation (ABC, Sisson et al., 2019) permettant d'estimer directement la distribution des paramètres. Nous discutons les avantages et limites de chacune de ces deux méthodes dans le cadre de l'estimation des paramètres en jeu dans la dynamique de l'adventice (Section 4).

2 Extension du cadre HSMM au cas où la chaîne cachée est pilotée par les observations

2.1 Rappel sur les HSMM

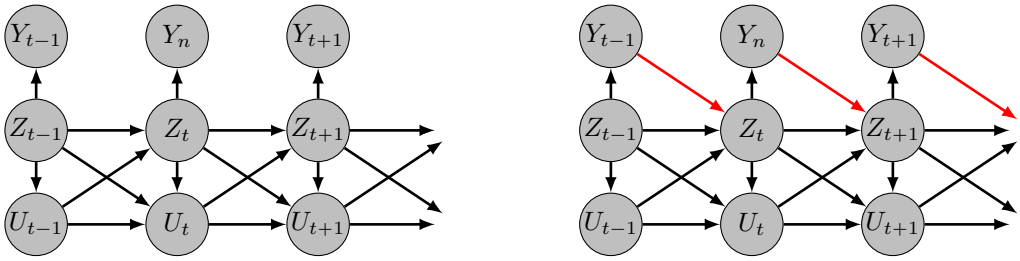
Nous considérons ici le cas temps discret et espaces d'états discrets. Un modèle de semi-Markov caché (Hidden Semi-Markov Model, HSMM) se compose de deux processus aléatoires, indexés par le temps : le processus observé $Y = (Y_t)_{t\in\mathbb{N}}$, dont l'espace d'états est $\Omega_Y = \{1, ..., D\}$, et le processus caché $Z = (Z_t)_{t\in\mathbb{N}}$, qui est une chaîne de semi-Markov avec pour espace d'états $\Omega_Z = \{1, ..., S\}$. Les observations entre $t = 0$ et $t = M$ sont désignées par le vecteur $Y_{0:M} = (Y_0, Y_1, Y_2, ..., Y_M)$. De même, le vecteur des états cachés entre $t = 0$ et $t = M$ est noté $Z_{0:M} = (Z_0, Z_1, ..., Z_M)$. Ainsi, dans la version classique du HSMM, la distribution jointe de $(Z_{0:M}, Y_{0:M})$ est entièrement déterminée par les distributions suivantes : la probabilité initiale $\mathbb{P}(Z_0 = i)$ (notée $\pi(i)$), la probabilité d'émission $\mathbb{P}(Y_t = a|Z_t = i)$ (notée $R(i, a)$) et enfin, le noyau de transition $\mathbb{P}(Z_{t+d} = j, Z_{t+1:t+d-1} = i|Z_t = i, Z_{t-1} \neq i)$ (noté $q_{ij}(d)$), où d est la durée de séjour de la chaîne cachée dans l'état i. Le noyau de transition du HSMM est le produit de la probabilité que la durée de séjour de la chaîne cachée dans un état i soit d, sachant que l'état suivant sera j, et de la probabilité de transition de cet état vers l'état j.

2.2 Comment introduire une dépendance aux observations dans un HSMM ?

Dans ce travail, on cherche à étendre le HSMM classique au cas où l'observation Y_{t-1} a une influence sur la variable cachée suivante Z_t. Dans ce cas, on ne peut plus définir une variable aléatoire représentant la durée de séjour conditionnellement à l'entrée dans l'état puisqu'à chaque instant l'impact de l'observation peut venir modifier cette durée. La définition d'un HSMM à partir du noyau $q_{ij}(d)$ ne semble donc pas adaptée. Il existe une autre expression d'un HSMM, basée sur l'introduction d' un nouveau processus $U = (U_t)_{t\in\mathbb{N}}$ décrivant le temps passé depuis l'entrée dans l'état Z_t, avec comme espace d'états $\Omega_U = \mathbb{N}^*$. Ainsi, alors que Z_t est une chaîne de semi-Markov, le couple (Z_t, U_t) forme une chaîne de Markov (Barbu and Limnios, 2008). Cette modélisation d'un HSMM est donc plus adaptée pour une extension au cas où les observations (Y_t) ont une influence, à chaque pas de temps, sur les états cachés (Z_t, U_t). C'est celle que nous adoptons.

2.3 Définition générale du modèle OD-HSMM

Le modèle OD-HSMM (Observation-Driven HSMM) est composé de trois processus aléatoires : le processus observé $Y = (Y_t)_{t\in\mathbb{N}}$, le processus caché $Z = (Z_t)_{t\in\mathbb{N}}$ et le processus $U = (U_t)_{t\in\mathbb{N}}$ (caché également) des temps passés depuis l'entrée dans le dernier état. Soit $U_{0:M} = (U_0, U_1, U_2, ..., U_M)$. Dans un OD-HSMM la distribution jointe de $(Z_{0:M}, U_{0:M}, Y_{0:M})$ est définie à partir des distributions suivantes :



(a) HSMM (b) OD-HSMM

FIGURE 1 – Représentation graphique des dépendances conditionnelles dans la chaîne (Z_t, U_t, Y_t) : (a) cadre HSMM, (b) cadre OD-HSMM.

— La loi initiale $\mathbb{P}(Z_0 = z_0, U_0 = 1) = \mathbb{P}(Z_0 = z_0)$ est notée $\pi(z_0)$ (on suppose que l'on entre dans un nouvel état à $t = 0$) ;
— La probabilité d'émission (qui reste indépendante de U_t comme dans un HSMM) $\mathbb{P}(Y_t = y_t | Z_t = z_t, U_t = u_t) = \mathbb{P}(Y_t = y_t | Z_t = z_t)$ est notée $R(z_t, y_t)$;
— La probabilité de transition $\mathbb{P}(Z_t = z_t, U_t = u_t | Z_{t-1} = z_{t-1}, U_{t-1} = u_{t-1}, Y_{t-1} = y_{t-1})$ est notée $P_{y_{t-1}}((z_{t-1}, u_{t-1}), (z_t, u_t))$.

Les dépendances conditionnelles dans un OD-HSMM sont représentées dans la Figure 1.

Remarque 1. *En raison de l'influence des observations sur les états cachés, la probabilité de transition dépend de l'observation y_{t-1}. Donc pour $z_t, u_t, z_{t-1}, u_{t-1}$ fixés il y a autant de probabilité de transition à calculer que d'états observés possibles.*

Remarque 2. *La probabilité de transition est à définir quelque soit $z_t, u_t, z_{t-1}, u_{t-1}$. Cependant, seul un faible nombre de probabilités sont non nulles car u_t ne peut valoir que $u_{t-1} + 1$ ou 1.*

3 Modélisation de la dynamique de l'adventice avec un OD-HSMM paramétrique

3.1 Paramètres en jeu dans la dynamique de l'adventice

L'adventice est une plante annuelle, c'est-à-dire qu'elle meurt chaque année. Les paramètres en jeu dans la dynamique d'une population d'adventices sont la germination, la grenaison, la colonisation et la survie de la banque de graines. En s'inspirant de la modélisation de type OD-HMM avec espaces d'états en présence/absence proposée par Pluntz et al. (2018), nous définissons les paramètres suivants à partir desquels nous allons construire les probabilités d'émission et de transition du OD-HSMM.

1. Une partie des graines dans la banque de graines peut germer et produire des plantes, avec probabilité g.
2. Indépendamment d'une possible germination, le stock de graines peut survivre dans le sol, mais cette survie diminue avec le temps passé dans le sol u. On définit donc la

probabilité que la banque de graines survive entre t et $t+1$ si elle existe depuis déjà u pas de temps par

$$s = s_0 e^{-\lambda(u-1)},$$

où s_0 est la probabilité qu'un nouveau stock de graines créé à t survive à $t+1$ et λ modélise la vitesse avec laquelle la banque de graine s'épuise.

3. Une fois que les graines ont poussé et sont devenues des plantes en fleur, ces dernières dispersent leurs graines au sol de la parcelle. Celles-ci entrent donc dans la banque de graines. C'est ce qu'on appelle la grenaison, qui est de probabilité d.
4. Chaque année, la quantité de graines dans le sol peut augmenter grâce à la colonisation par des graines provenant d'un autre champ. On suppose une colonisation de probabilité constante, c.

On notera p_0 la probabilité qu'il y ait déjà des graines dans le sol la première année d'observation.

3.2 Lois du modèle

Comme dans Pluntz et al. (2018), nous nous plaçons dans le cadre présence/absence (i.e. avec $\Omega_Z = \Omega_Y = \{0,1\}$). Soit $\theta = (p_0, c, s_0, d, g, \lambda)$ le vecteur des paramètres, les lois de l'OD-HSMM s'écrivent de la façon suivante en fonction de θ.

La loi initiale $\pi(z_0)$ s'exprime en fonction de p_0 : $(\pi(0), \pi(1)) = (1-p_0, p_0)$.

La loi d'émission $R(z_t, y_t)$ représente la germination des graines d'adventices. Elle s'exprime donc en fonction de g, de la façon suivante :

$$R = \begin{pmatrix} 1 & 0 \\ 1-g & g \end{pmatrix},$$

où le premier élément de la première ligne désigne la probabilité $\mathbb{P}(Y_t = 0 | Z_t = 0)$, qui vaut 1 puisqu'il ne peut pas y avoir de plante en fleur lorsqu'il n'y a pas de graine dans le sol.

La probabilité de transition dépend de u_{t-1} et u_t qui ne sont pas bornés. Pour faciliter la présentation, nous présentons les différentes valeurs prises par $P_{y_{t-1}}((z_{t-1}, u_{t-1}), (z_t, u_t))$ via une « matrice » de transition décomposée en 4 blocs selon les valeurs de z_{t-1} et z_t même si ces blocs sont de taille infinie ($|\Omega_U| \times |\Omega_U|$). Ces 4 blocs correspondent au 4 états possibles du couple (Z_t, U_t), ils sont notés $A^k_{i,j}$, où $\forall u_{t-1}, u_t,\ A^k_{i,j}(u_{t-1}, u_t) = P_k(i, u_{t-1}, j, u_t)$. On exprime les deux « matrices » de transition en fonction des $A^k_{i,j}$ comme suit :

$$P_0 = \left(\begin{array}{c|c} A^0_{0,0} & A^0_{0,1} \\ \hline A^0_{1,0} & A^0_{0,1} \end{array}\right)$$

$$P_1 = \left(\begin{array}{c|c} A^1_{0,0} & A^1_{0,1} \\ \hline A^1_{1,0} & A^1_{0,1} \end{array}\right)$$

Avec :

$$A_{ij}^k = \begin{cases} \begin{pmatrix} 0 & \alpha_i^k(1) & 0 & \dots \\ 0 & 0 & \alpha_i^k(2) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} & \text{si } i = j, \\ \begin{pmatrix} \beta_i^k(1) & 0 & 0 & \dots \\ \beta_i^k(2) & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix} & \text{si } i \neq j. \end{cases}$$

Où $\forall k \in \{0,1\}$:

$$\alpha_i^0(u) = \begin{cases} (1-c) \text{ si } i = 0 \\ 1-(1-c)(1-s_0 \times e^{-\lambda(u-1)}) \text{ si } i = 1 \end{cases},$$

$$\alpha_i^1(u) = \begin{cases} (1-c)(1-d) \text{ si } i = 0 \\ 1-(1-c)(1-s_0 \times e^{-\lambda(u-1)})(1-d) \text{ si } i = 1 \end{cases},$$

et $\forall k \in \{0,1\},\ \beta_i^k(u) = 1 - \alpha_i^k(u)$.

Example 1 (Détail du calcul de $\alpha_1^0(2)$)**.**

$$\begin{aligned} \alpha_1^0(2) = & \quad \mathbb{P}(Z_t = 1, U_t = 3 | Z_{t-1} = 1, U_{t-1} = 2, Y_{t-1} = 0) \\ = & \quad 1 - \mathbb{P}(Z_t = 0, U_t = 1 | Z_{t-1} = 1, U_{t-1} = 2, Y_{t-1} = 0), \end{aligned}$$

où la probabilité $\mathbb{P}(Z_t = 0, U_t = 1 | Z_{t-1} = 1, U_{t-1} = 2, Y_{t-1} = 0)$ *décrit l'événement dans lequel il n'y a plus de graine dans le sol alors qu'il y en avait précédemment depuis deux pas de temps mais qu'il n'y avait pas de flore levée. Ainsi, les graines dans le sol n'ont pas survécu, événement dont la probabilité vaut* $(1 - s_0 \times e^{-\lambda(2-1)})$ *et il n'y a pas non plus eu de colonisation, ce qui est de probabilité* $(1-c)$.

3.3 Exemple

Afin de rendre plus compréhensible le lien entre les paramètres et les probabilités de transition, nous présentons ici les 3 premières lignes et colonnes des blocs $A_{i,j}^k$. Pour alléger

les notations, on pose $s_u = s_0 \times e^{-\lambda(u-1)}$.

$$P_0 = \left(\begin{array}{cccc|cccc} 0 & 1-c & 0 & \dots & c & 0 & 0 & \dots \\ 0 & 0 & 1-c & \dots & c & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots & c & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \hline (1-c)(1-s_1) & 0 & 0 & \dots & 0 & 1-(1-c)(1-s_1) & 0 & \dots \\ (1-c)(1-s_2) & 0 & 0 & \dots & 0 & 0 & 1-(1-c)(1-s_2) & \dots \\ (1-c)(1-s_3) & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ \dots & & \dots & \dots & \dots & \dots & \dots & \dots \end{array}\right)$$

$$P_1 = \left(\begin{array}{cccc|cccc} 0 & (1-c)(1-d) & 0 & \dots & 1-(1-c)(1-d) & 0 & 0 & \dots \\ 0 & 0 & (1-c)(1-d) & \dots & 1-(1-c)(1-d) & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots & 1-(1-c)(1-d) & 0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \hline (1-c)(1-s_1)(1-d) & 0 & 0 & \dots & 0 & 1-(1-c)(1-s_1)(1-d) & 0 & \dots \\ (1-c)(1-s_2)(1-d) & 0 & 0 & \dots & 0 & 0 & 1-(1-c)(1-s_2)(1-d) & \dots \\ (1-c)(1-s_3)(1-d) & 0 & 0 & \dots & 0 & 0 & 0 & \dots \\ & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{array}\right)$$

4 Estimation des paramètres de l'OD-HSMM pour adventice

Nous explorons deux approches : l'approche EM, naturelle dans le cas d'un modèle avec données manquantes ainsi que l'approche ABC, adaptée au cas où la vraisemblance n'est pas accessible.

4.1 Approche EM

L'algorithme EM repose sur la quantité intermédiaire suivante, où $\theta^{(m)}$ est la valeur courante du paramètre :

$$Q(\theta|\theta^{(m)}) = \mathbb{E}\left[\ln \mathbb{P}(Y_{0:M}, Z_{0:M}, U_{0:M}|\theta)|Y_{0:M} = y_{0:M}, \theta^{(m)}\right]$$

Il s'agit d'un algorithme itératif où chaque itération est composée de deux étapes : à l'étape E on calcule les distributions marginales intervenant dans l'expression de la quantité intermédiaire, puis à l'étape M on met à jour l'ensemble des paramètres en résolvant $\theta^{(m+1)} = \arg\max_{\theta} Q(\theta|\theta^{(m)})$.

Dans le cas de l'OD-HSMM pour estimer la dynamique de l'adventice, il n'existe pas de formule explicite pour mettre à jour les paramètres c, d, s_0 et λ lors de l'étape M. Pour palier à ça, il est possible de résoudre le problème de maximisation de façon numérique en utilisant, par exemple, des outils disponibles dans R. Quelques essais préliminaires n'ont pas conduit à des résultats satisfaisants, ainsi nous explorons une autre approche. Dans le cas non paramétrique, l'étape M de mise à jour des probabilités de transition et d'émission (P_y, R)

conduit à des expressions analytiques simples. Nous choisissons donc de mettre en oeuvre l'EM sur l'OD-HSMM non paramétrique et d'ajouter une étape supplémentaire d'identification des paramètres θ à partir des P_y et R estimés, à l'issue du EM.

4.1.1 Étape E

L'étape E de l'algorithme EM pour l'OD-HSMM est très similaire à celle de l'algorithme EM pour HMM (mais avec une variable cachée bi-dimensionnelle) et repose sur l'algorithme Forward-Backward. Cependant, la récursion Backward a été modifiée pour prendre en compte le fait que la probabilité de transition dépend de l'observation.
A priori l'expression de $Q(\theta|\theta^{(m)})$ faut intervenir des sommes infinies sur U_t et U_{t-1} mais en pratique ces variables sont bornées par M puisque l'on suppose qu'à $t = 0$ la chaîne cachée entre dans un nouvel état.

4.1.2 Étape M

L'étape M consiste à mettre à jour l'ensemble des probabilités d'émission et de transition, noté $\theta^{NP} = (\{P_y((i,u),(j,u'))\}, \{R(i,a)\})$. Pour cela, on résout le problème de maximisation $\arg\max_{\theta^{NP}} Q(\theta^{NP}|\theta^{NP(m)})$. Les formules de mise à jour sont explicites. Par exemple pour $R(1,1)$ on obtient

$$R(1,1)^{(m+1)} = \frac{\sum_{t=0}^{M} \mathbb{P}(Z_t = 1|Y_{0:M} = y_{0:M})\mathbb{1}_{y_t=1}}{\sum_{t=0}^{M} \mathbb{P}(Z_t = 1|Y_{0:M} = y_{0:M})}.$$

4.1.3 Identification des paramètres

Le paramètre g étant égal à $R(1,1)$ son estimation est directe à partir des estimateurs $\hat{\theta}^{NP} = (\{\hat{P}_y((i,u),(j,u'))\}, \{\hat{R}(i,a)\})$ obtenus à l'issue du EM. Soit $P_y(\theta)$ l'expression des probabilités de transition en fonction des paramètres (c, s_0, d, λ) du modèle adventice (cf Section 3). L'étape d'identification des paramètres consiste à minimiser la distance entre $\hat{P}_y$ et $P_y(\theta)$, grâce à des outils classiques de résolution de systèmes non linéaires.

4.2 Approche ABC

L'algorithme ABC est un algorithme d'estimation bayésienne qui consiste à générer un grand nombre de valeurs pour les paramètres θ du modèle et à identifier ceux qui conduisent à des données simulées « proches » de celles observées. En pratique, pour une valeur échantillonnée $\hat{\theta}$, la première étape consiste à générer des données simulées selon le modèle. Puis, dans une seconde étape des statistiques sont calculées sur ces données simulées et comparées à celles calculées sur les observations réelles. Dans le cas où la distance entre

les deux est inférieure à un certain seuil, les paramètres $\hat{\theta}$ utilisés pour la simulation sont acceptés. L'ensemble des valeurs $\hat{\theta}$ retenues fournit une distribution a posteriori de θ.

Pour mettre en oeuvre ABC pour estimer les paramètres de l'OD-HSMM adventice, nous utilisons le package R « EasyABC » (Jabot et al., 2013). Nous proposons d'utiliser les statistiques suivantes calculées sur $y_{0:M}$:

— le nombre de 0 ;
— le minimum, le maximum, la moyenne et les quantiles à 25%, 50% et 75% des longueurs des phases de 1 consécutifs et des phases de 0 consécutifs de 0 et 1 ;
— le nombre de transitions de 0 vers 1.

Un avantage à utiliser l'algorithme ABC est que cette méthode ne nécessite pas de savoir calculer la vraisemblance, elle repose uniquement sur la simulation du modèle, qui est très simple pour le modèle OD-HSMM. Enfin, l'algorithme ABC est une méthode d'estimation bayésienne donc il permet d'obtenir une estimation de la distribution des paramètres plutôt qu'un estimateur ponctuel comme avec l'algorithme EM. En revanche, choisir les bonnes statistiques peut s'avérer être un exercice compliqué. C'est ce que nous sommes en train d'explorer sur des données simulées.

5 Conclusion

Nous proposons un modèle général, l'OD-HSMM, permettant de combiner les avantages du cadre HSMM et du cadre OD-HMM. Ce modèle est motivé par une application à la modélisation de la dynamique des adventices mais il peut avoir aussi un intérêt dans d'autres domaines comme la finance (Engel and Hamilton, 1990) pour prédire le régime d'un système monétaire en fonction du taux de change.

Nous décrivons ensuite une version paramétrique de l'OD-HSMM, pour la dynamique d'une adventice. Nous comparons deux méthodes d'estimation. La première consiste à estimer les lois du modèle général non paramétrique grâce à l'algorithme EM, puis à identifier les paramètres adventice par résolution d'un système non linéaire. La seconde estime la distribution des paramètres par le biais de l'algorithme ABC. A ce stade ABC nous semble plus adapté du fait de la simplicité de mise en oeuvre, mais des expérimentations sur données simulées sont en cours afin de comparer les performances des algorithmes EM et ABC. Ensuite, l'algorithme le plus performant sera appliqué à un jeu de données réel sur les adventices (Pluntz et al., 2018).

Une perspective à ce travail est d'étendre le modèle à un cadre spatial plus réaliste en modélisant explicitement la colonisation entre parcelles au lieu de supposer qu'il s'agit d'une pluie de graines. Pour cela, nous explorerons l'extension de l'OD-HSMM au cas multi-chaînes où chaque chaîne décrit la dynamique de l'adventice dans une parcelle donnée. La modélisation de l'influence de la flore levée d'une parcelle sur la banque de graines d'une autre pose les mêmes difficultés de modélisation que l'OD-HSMM, à savoir qu'une variable extérieure vient influer sur le temps de séjour de la chaîne locale. Une piste consistera donc à ré-exploiter la formulation d'un HSMM par le couple (Z, U). L'extension de l'OD-HSMM

au cas multi-chaînes pourra être comparée à la modélisation de Le Coz et al. (2019) qui est moins générale car dans le cadre HMM et non HSMM.

Références

Bacave, H., P.-O. Cheptou, N. Limnios, and N. Peyrard (2023). Non parametric Observation Driven HMM. Preprint available at https://hal.inrae.fr/hal-04053732v1.

Barbu, V.-S. and N. Limnios (2008). *Semi-Markov Chains and Hidden Semi-Markov Models towards Applications*. Springer.

Baskin, C. and J. Baskin (2014). *Seeds : Ecology, Biogeography, and, Evolution of Dormancy and Germination*. San diego, Academic Press.

Dempster, A., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*, 1–22.

Engel, C. and J.-D. Hamilton (1990). Long swings in the dollar : Are they in the data and do markets know it ? *The American Economic Review 80*(4), 689–713.

Jabot, F., T. Faure, and N. Dumoulin (2013). EasyABC : performing efficient approximate Bayesian computation sampling schemes using R. *Methods in Ecology and Evolution*, 684–687.

Le Coz, S., P. O. Cheptou, and N. Peyrard (2019). A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy. *Theoretical Population Biology 127*, 120–132.

Pluntz, M., S. Le Coz, N. Peyrard, R. Pradel, R. Coquet, and P. O. Cheptou (2018). A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora. *Ecology Letters 21*, 1311–1318.

Sisson, A. S., F. Yanan, and M. A. Beaumont (2019). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC.

Yu, S.-Z. (2016). *Hidden Semi-Markov Models Theory, Algorithms and Applications*. Elsevier.