Un cadre HMM spatial pour modéliser la dynamique d'espèces avec stade de dormance

Nathalie Peyrard [1] & Sebastian Le Coz [2] & Pierre-Olivier Cheptou [3]

[1] *INRA UR 875 MIAT, Chemin de Borde Rouge, 31326 Castanet-Tolosan, France, nathalie.peyrard@inra.fr*

[2] *INRA UR 875 MIAT, Chemin de Borde Rouge, 31326 Castanet-Tolosan, France, sebastian.le-coz@outlook.fr*

[3] *CEFE UMR 5175,CNRS, Université de Montpellier, Université Paul-Valery Montpellier, EPHE - 1919, route de Mende - 34293 Montpellier cedex 05, France, pierre-olivier.cheptou@cefe.cnrs.fr*

Résumé. De nombreuses espèces ont un stade dormant dans leur cycle de vie, par exemple les graines pour les plantes. L'état de la population dormante influence la dynamique de l'espèce, cependant elle est souvent difficilement détectable. Une façon d'inclure la dormance dans un modèle dynamique est alors de considérer l'état de la population dormante comme un état caché et de se placer dans le cadre des modèles de Markov cachés. De tels modèles ont déjà été proposés mais avec plusieurs limites : les populations dormantes et non dormantes sont modélisées comme des variables binaires (présence-absence), la durée de la dormance est limitée à un seul pas de temps, la colonisation par les patchs voisins n'est pas prise en compte. Nous proposons un modèle de Markov caché qui lève ces limites et permet ainsi de mieux décrire à la fois la dynamique locale et régionale d'une espèce avec dormance : le modèle de Markov caché multidimensionnel avec rétroaction des données. Pour un modèle Markovien multidimensionnel, la complexité de l'estimation des paramètres du modèle par l'algorithme EM est a priori exponentielle en fonction du nombre de chaînes, ainsi que la complexité de restauration de l'état caché et de la prédiction. Cependant nous démontrons que pour le modèle proposé ces requêtes sont réalisables pour une complexité linéaire. Des tests sur des données simulées montrent que les estimateurs obtenus sont de bonne qualité, ainsi que les performances en terme de restauration et de prédiction. Ce nouveau cadre fournit un outil simple et efficace pour l'analyse et la comparaison des dynamiques de plantes, comme par exemple les espèces adventices dans les parcelles cultuvées.

Mots-clés. Métapopulation, Banque de graines, Stade non observé, Modèle de Markov caché, algorithme EM

Abstract. Many species have a dormant stage in their life cycle, including seeds for plants. The dormancy stage influences the species dynamics but is often undetectable. One way to include dormancy is to model it as a hidden state in a hidden Markov model. Models within this framework have already been proposed but with different limitations: only presence/absence observations are modelled, the dormancy stage is limited to one

year, or colonisation from neighbouring patches is not taken into account. We propose a hidden Markov model that does not have these limits and which enables a better representation of both the local and regional dynamics of a species with dormancy: the multidimensional HMM with data feedback. Parameter estimation for multidimensional HMM using EM a priori has an exponential computational time in terms of the number of chains, as well as restoration and prediction. However, we demonstrate that for our model structure, these tasks are achievable in a linear computational time. Experiments on simulated data show that estimation is of good quality, as well as restoration and prediction. This framework provides a simple and efficient tool that could be further used to analyse and compare annual plants dynamics like weed species in crop fields.

Keywords. Metapopulation, Seed bank, Hidden life stage, Hidden Markov model, EM algorithm

1 Introduction

De nombreuses espèces ont un stade dormant dans leur cycle de vie, par exemple les graines pour les plantes. L'état de la population dormante influence la dynamique de l'espèce, cependant elle est souvent difficilement détectable. Une façon d'inclure la dormance dans un modèle dynamique est alors de considérer l'état de la population dormante comme un état caché et de se placer dans le cadre des modèles de Markov cachés (Hidden Markov Models, HMM). De tels modèles ont déjà été proposés (e.g. **??**) mais avec plusieurs limites : les populations dormantes et non dormantes sont modélisées comme des variables binaires (présence-absence), la durée de la dormance est limitée à un seul pas de temps, la colonisation par les patchs voisins n'est pas prise en compte. Nous proposons un HMM qui lève ces limites et permet ainsi de mieux décrire à la fois la dynamique locale et régionale d'une espèce avec dormance (Section 2) : le modèle de Markov caché multidimensionnel avec rétroaction des données (Multidimensional HMM with Data Feedback, MHMM-DF). La rétroaction des données correspond au fait que dans ce modèle il y a une dépendance directe de l'état caché à l'état observé au même pas de temps. Cela correspond par exemple à la production de nouvelles graines par les plantes adultes, qui vont ensuite alimenter la banque de graine locale ou dans le voisinage.

L'algorithme classique pour estimer les paramètres d'un modèle avec données cachées est l'algorithme EM (**?**). Dans le cas des HMM, l'étape E correspond à l'algorithme Forward-Backward (**?**), qui s'appuie sur la structure linéaire d'un HMM et l'élimination de variable pour calculer efficacement toutes les probabilités conditionnelles des variables cachées sachant les observations. Dans le cas où la variable cachée est multidimensionnelle et qu'il y a donc plusieurs chaînes cachées en interaction, la taille des domaines et la structure des interactions peuvent être rédhibitoires pour le Forward-Backward. Cependant, dans un MHMM-DF les chaines cachées sont indépendantes conditionnellement

aux observations. Cela nous permet de proposer une version du Forward-Backward qui reste exacte et qui s'applique chaine par chaine indépendamment, pour une complexité en temps qui est linéraire en le nombre de chaînes, au lieu d'exponentielle (Section 3). Nous présentons en Section 4 des résultats sur le comportement de l'algorithme EM associé, puis des résultats en terme de sélection de modèle, restauration et prédiction, obtenus sur des données simulées.

2 Modèle

Considérons un ensemble de C patchs. Aux temps $n \in \{1, \ldots, N\}$ sur le patch c deux variables sont définies : $X_{c,n}$ est la classe d'abondance de la population dormante (variable non observée, par exemple l'état de la banque de graines pour les plantes) et $Y_{c,n}$ est la classe d'abondance de la population non dormante (variable observée, par exemple l'abondance de flore levée pour les plantes). Leurs domaines sont $\Omega_X = \{0, 1, \ldots, |\Omega_X|-1\}$ et $\Omega_Y = \{0, 1, \ldots, |\Omega_Y| - 1\}$, respectivement. Un MHMM-DF de dimension C modélise la dynamique jointe des C chaines sous deux hypothèses. Tout d'abord, pour un patch c au pas de temps n, l'état de la population non dormante $Y_{c,n+1}$ ne dépend que de l'état de la population dormante locale au pas de temps précédent, $X_{c,n}$. Ensuite, l'état de la population dormante $X_{c,n+1}$ dépend de trois processus via trois (ensembles de) variables : (1) la survie de la population dormante, via $X_{c,n}$, (2) la production de nouveaux individus dormants, via $Y_{c,n+1}$ (dynamique locale) et (3) de l'arrivée par colonisation de nouveaux individus dormants, via les états des populations non dormantes des patchs pouvant coloniser à $n + 1$ (dynamique spatiale). La dépendance directe orientée de $Y_{c,n+1}$ vers $X_{c,n+1}$ constitue la partie 'rétroaction des données' du modèle. La structure de dépendance est illustrée sur la figure 1 dans le cas de deux patchs.

Un MHMM-DF est défini par trois probabilités. La première est la probabilité initiale des états cachés, $\pi(x_{c,0})$. Ensuite, la probabilité d'émission, $\phi(x_{c,n-1}, y_{c,n}) = \mathbb{P}(Y_{c,n} = y_{c,n}|X_{c,n-1} = x_{c,n-1})$, modélise le processus de 'réveil', ce qui correspond dans le cas des plantes à la germination des graines et à la croissance des plantes jusqu'à l'âge adulte. Enfin, la probabilité de transition de la variable cachée de la chaîne c, définie comme $A(x_{c,n-1}, x_{c,n}, y_n^C) = \mathbb{P}(X_{c,n} = x_{c,n}|X_{c,n-1} = x_{c,n-1}, Y_n^C = y_n^C)$, où $Y_n^C = \{Y_{c,n}\}_{1 \leq c \leq C}$, modélise dans le cas des plantes les processus de survie des graines, de production locale de nouvelles graines et d'arrivée de nouvelles graines par dispersion spatiale.

Nous considérons le cas où le nombre de données disponible est faible, et nous nous plaçons dans un cadre paramétrique pour modéliser ϕ et A. Ces fonctions sont modélisées comme des régressions binomiales, où la probabilité de succès est exprimée en fonction des variables explicatives via une régression logistique. Ainsi, $\phi(x_{c,n-1}, y_{c,n})$ est une distribution binomiale de paramètres $|\Omega_Y|$ et $p_{x_{c,n-1}} = \frac{1}{1+\exp(-(\mu_0+\mu_1 x_{c,n-1}/|\Omega_X|))}$, avec μ_0 et μ_1 deux hyperparameters à estimer. Dans le cas de A, la probabilité de succès, $p_{x_{c,n-1},y_n^C}$ est également modélisée par une régression logistique dont les variables explicatives sont

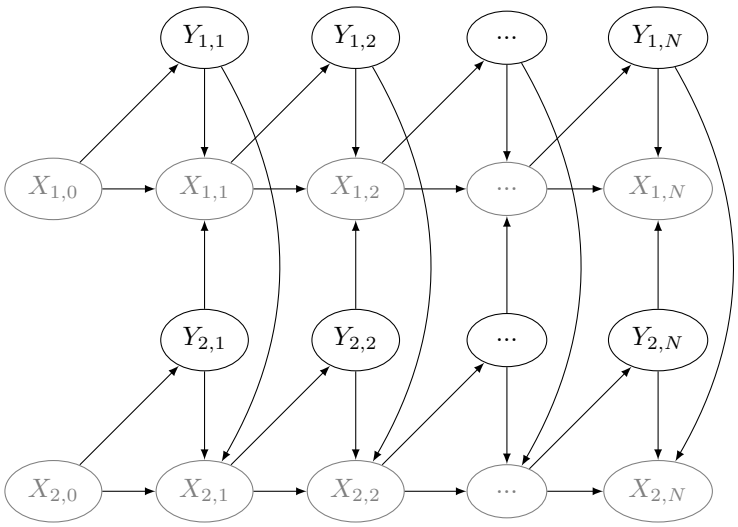

Figure 1: Graphe de dépendance d'un MHMM-DF, dans le cas de deux chaînes cachées (les nœuds gris représentent les variables cachées).

$x_{c,n-1}$, $y_{c,n}$ et $y_n^{\mathcal{C}\setminus c} = y_n^C \setminus y_{c,n}$ (la dernière variable pouvant être remplacé par les observations sur un sous-ensemble seulement des patchs, les voisins par exemple). Afin d'illustrer le modèle, nous avons considéré une modélisation de A où tous les patchs autres que le patch local c contribuent à la colonisation de c. Leur influence est mesurée par l'état moyen sur l'ensemble de ces patchs, qui donne une mesure de la capacité de colonisation moyenne. Enfin, $\pi(x_{c,0})$ est une loi binomiale de paramètres $|\Omega_X|$ et p_τ. Avec ces choix de paramétrisation, un MHMM-DF dépend de 7 paramètres. Dans ce cas particulier, nous avons démontré l'identifiabilité générique dès que $N > 7$, dès lors que $C > 2$.

3 Estimation

Nous présentons ici une mise en œuvre de l'algorithme EM pour MHMM-DF pour lequel l'étape E est de complexité linéaire en fonction du nombre de chaînes. Ce résultat est indépendant d'un choix de paramétrisation de π, A et ϕ. Notons $X^{c,n} = \{X_{c',n'}, 1 \leq c' \leq c, 1 \leq n' \leq n\}$ et $Y^{c',n'} = \{Y_{c',n'}, 1 \leq c' \leq c, 1 \leq n' \leq n\}$ Alors, soit $\lambda = (\pi, \phi, A)$ le vecteur des paramètres du modèle, nous définissons

$$Q(\lambda \mid \lambda') = E[\ln(\mathbb{P}(X^{C,N}, Y^{C,N}|\lambda)|Y^{C,N} = y^{C,N}, \lambda')].$$

L'algorithme EM itère sur les deux étapes suivantes :
Etape E : calcul des probabilités conditionnelles intervenant dans $Q(\lambda \mid \lambda_t)$ pour la valeur courante de l'estimateur λ_t
Etape M : mise à jour de λ par maximisation de $Q(\lambda \mid \lambda_t)$.

Dans le cas d'un MHMM-DF, les C chaînes cachées sont indépendantes conditionnellement aux observations. Cela implique que la fonction $Q(\lambda \mid \lambda')$ ne dépend que des probabilités conditionnelles suivantes.

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

Pour résoudre l'étape E, il est possible d'écrire un algorithme de Forward-Backward par chaîne, dont les variables auxiliaires pour la chaîne c sont

$$\alpha_{c,n}(x_{c,n}) = \mathbb{P}(Y^{C,n} = y^{C,n}, X_{c,n} = x_{c,n} | \lambda_t),$$

$$\beta_{c,n}(x_{c,n}) = \mathbb{P}(Y^{C}_{n+1} = y^{C}_{n+1}, ..., Y^{C}_{N} = y^{C}_{N} | Y^{C,n} = y^{C,n}, X_{c,n} = x_{c,n}, \lambda_t).$$

Le fait d'introduire $Y^{C,n} = y^{C,n}$ dans le conditionnement de $\beta_{c,n}(x_{c,n})$ permet de préserver la propriété $\alpha_{c,n}(x_{c,n})\beta_{c,n}(x_{c,n}) = \mathbb{P}(X_{c,n} = x_{c,n}, y^{C,N} | \lambda_t)$, nécessaire au Forward-Backward. Ces variables auxiliaires peuvent être calculées de manières récursives, comme dans un Forward-Backward classique.

L'étape M est elle spécifique à la paramétrisation choisie. Elle peut être résolue par des approches numériques classiques.

4 Evaluation sur données simulées

Tous les résultats ont été obtenus à partir de données simulées selon le modèle MHMM-DF à 7 paramètres décrit précédemment, pour différentes valeurs des paramètres. Les données ont été simulées pour un modèle à $C = 10$ patchs, $N = 100$ pas de temps et $|\Omega_X| = |\Omega_Y| = 5$. Nous avons tout d'abord testé la qualité des estimateurs obtenus par l'algorithme EM. Lorsque les valeurs des paramètres correspondent à des dynamiques extrêmes (absence presque dans tous les patchs, ou au contraite présence dans la classe maximale dans presque tous les patchs), l'estimation est difficile, sans surprise. Pour des dynamiques qui visitent tous les états, les estimateurs obtenus sont de bonne qualité (biais et variance). Nous avons ensuite testé la possibilité de distinguer entre une dynamique avec et sans colonisation spatiale, ou une dynamique avec ou sans survie de la population dormante. La sélection est faite suivant le critère AIC, pour des paramètres estimés par EM. Nous avons observé de meilleurs résultats sur la détection de l'absence de dormance que sur la détection d'absence de colonisation. Enfin, nous avons évalué la qualité de la prédiction de l'état des populations non dormantes au pas de temps suivant et de la restauration de l'état des populations dormantes. Pour cela nous avons adapté l'algorithme Viterbi pour une mise en œuvre chaîne par chaîne, comme pour EM. La qualité de la restauration est rarement en dessous de 70 %. La qualité de la prédiction peut atteindre 80 % pour certaines valeurs de paramètres mais peut chuter à 40% dans des situations où le mode est peu piqué. Du fait de la complexité linéaire du EM, les mêmes expériences ont pu être mises en œuvre dans le cas d'un MHMM-DF à 100 chaînes, pour des résultats similaires.

5 Conclusion

Le cadre MHMM-DF permet une modélisation de dynamique d'espèces avec stade de dormance tenant compte également de la colonisation spatiale. Du fait de la structure du modèle, l'estimation, la restauration ou encore le calcul de la vraisemblance sont accessibles, contrairement à d'autres cadres de HMM multidimensionnels. Une description étendue de ces travaux est présentée dans (**?**). Cette modélisation est particulièrement adaptée à l'étude de la dynamique de plantes annuelles. Chaque processus étant associé à un paramètre, il est facile d'identifier les processus prépondérants. Ainsi ces travaux sont en cours d'application sur des donnés d'espèces adventices. Les applications dépassent le cas des espèces annuelles car il est encore possible de définir un algorithme EM de complexité linéaire si l'on rajoute des arêtes d'observation à n vers observation à $n+1$ (survie des populations non dormantes d'un pas de temps sur l'autre) ou si on rajoute une arête d'une observation au patch c à n vers une observation au patch c' à $n+1$ (déplacement des populations non dormantes).

Remerciements. Ces travaux ont été en partie financés par le projet ANR AGROBIOSE (ANR-2013-0001) et par la Région Occitanie.

References

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Fréville, H., R. Choquet, R. Pradel, and P. Cheptou (2013). Inferring seed bank from hidden Markov models: new insights into metapopulation dynamics in plants. *Journal of Ecology 101*(6), 1572–1580.

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 (en révision)*.

Pluntz, M., S. L. Coz, N. Peyrard, R. Pradel, R. Choquet, and P.-O. Cheptou (2018). A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora. *Ecology Letters 21*(9), 1311–1318.

Rabiner, L. R. (1989, Feb). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.