Introduction

Nathalie PEYRARD[1], Stéphane ROBIN[2] et Olivier GIMENEZ[3]
[1] MIAT, INRA, Université de Toulouse, Castanet-Tolosan, France
[2] MIA Paris, INRAE, AgroParisTech, Université Paris-Saclay, Paris, France
[3] CEFE, CNRS, EPHE, IRD, Université Paul-Valéry Montpellier 3, Montpellier, France

I.1. Les variables cachées en écologie

L'écologie étudie les organismes vivants en interaction avec leur environnement. Ces interactions se produisent à l'échelle de l'individu (un animal, une plante), d'un groupe d'individus (une population, une espèce) ou de plusieurs espèces (une communauté). Pour étudier ces interactions, la statistique fournit les outils pour collecter, organiser, présenter, analyser, et tirer des conclusions à partir des données récoltées sur les systèmes écologiques. Il arrive toutefois que certaines composantes de ces systèmes écologiques échappent à l'observation, on parle alors de variables cachées. Cet ouvrage est consacré aux modèles intégrant des variables cachées en écologie et à leur inférence statistique.

Dans les différents chapitres, les variables cachées étudiées peuvent être rangées en trois classes correspondant à trois types de questions que l'on peut se poser sur un système écologique. On peut s'intéresser à identifier des groupes d'individus ou d'espèces. Il s'agit par exemple de groupes d'individus ayant un même comportement ou des profils génétiques proches, ou encore des groupes d'espèces qui interagissent avec les mêmes espèces, ou de la même manière à l'environnement. On peut aussi s'intéresser à étudier des variables dont on ne peut observer qu'une version bruitée,

Approches statistiques pour les variables cachées en écologie,
coordonné par Nathalie PEYRARD et Olivier GIMENEZ.

que l'on appelle souvent un « proxy ». On s'interrogera par exemple sur la présence d'une espèce que l'on peut manquer du fait d'une difficulté de détection ou d'erreurs de détection (confusion avec une autre espèce), ou encore de mesures bruitées du fait d'une erreur de mesure liée à la technologie. Enfin, on peut vouloir explorer les données, et réduire la dimension de l'information qu'elles contiennent, en se ramenant à un petit nombre de variables explicatives. Nous pouvons remarquer qu'entre le premier et le dernier cas, on s'éloigne de l'idée de variable qui échappe à l'observation et l'on généralise la notion de variable cachée.

Ces trois problèmes peuvent tous se traduire par des questions d'inférence sur des variables dites latentes en statistique. Cette inférence pose des problèmes statistiques qui requièrent des méthodes spécifiques que nous développerons. Nous nous attarderons également sur l'interprétation écologique de ces variables. Nous verrons que, même si leur traitement statistique est parfois complexe, leur considération dans les modèles permet de gagner en compréhension des systèmes écologiques.

I.2. Variable cachée et modélisation statistique

Le terme de variable cachée, que l'on utilise naturellement en écologie, se traduit en modélisation statistique par la notion plus générale de variable latente. Cette notion englobe plusieurs situations et va au-delà de la seule notion de variable physique non observable. En statistique, l'acception générale de la notion de variable latente est une variable d'intérêt qui n'est pas observable et n'a pas nécessairement de sens physique, et dont on va déduire la valeur à partir des observations. Si l'on essaie de préciser, les variables latentes sont caractérisées par les deux spécificités suivantes :

i) elles sont en nombre comparable au nombre de données, contrairement aux paramètres qui sont peu nombreux. Pensons par exemple à une chaîne de Markov cachée où le nombre de variables observées et de variables latentes est égal au nombre de pas de temps d'observation ;

ii) si leur valeur était connue alors l'estimation du modèle serait plus aisée. Pensons ici à l'estimation des paramètres d'un modèle de mélange lorsque les groupes des individus sont connus.

En pratique, si la variable latente a une réalité physique mais que l'on ne peut pas l'observer sur le terrain (la trajectoire exacte d'un animal, l'abondance de la banque de graines), il est commun de l'appeler variable cachée (néanmoins les deux termes sont souvent utilisés l'un pour l'autre). Dans d'autres cas, la variable latente intervient naturellement dans la description du processus ou du système étudié mais n'a pas d'existence physique. C'est ainsi le cas des variables latentes correspondant à l'une des observations en groupes. Nous parlerons alors de variables fictives. Enfin, les variables latentes peuvent également remplir une fonction instrumentale afin de

décrire une source de variabilité dans les observations non expliquées par les covariables connues, ou de décrire de manière synthétique une structure de dépendance. Elles peuvent être le résultat d'une méthode de réduction de dimension dans un groupe de variables explicatives dans un contexte de régression. On aura à l'esprit par exemple les composantes principales d'une analyse en composantes principales.

La notion de variable latente n'est pas sans lien avec la notion de modèle hiérarchique : les éléments des niveaux supérieurs du modèle, lorsqu'ils ne sont pas des paramètres du modèle, sont alors des variables latentes. Précisons également que la notion de variable latente peut être étendue au cas de quantités déterministes (c'est-à-dire que dans le modèle, c'est une constante). C'est le cas par exemple lorsque la variable latente est la trajectoire d'une équation différentielle ordinaire (EDO) pour laquelle on ne dispose que d'une observation bruitée.

I.3. Les outils offerts par la statistique

I.3.1. *Quelques modèles à variables latentes*

La liste des modèles statistiques impliquant des variables latentes est trop longue pour être faite ici. On peut cependant mentionner les modèles suivants, qui sont parmi les plus communs.

Les modèles de mélange permettent de définir un petit nombre de groupes dans lesquels se répartissent un ensemble d'observations. Les variables latentes sont alors des variables discrètes indiquant l'appartenance de chaque observation à l'un des groupes. Les modèles à blocs stochastiques (SBM) ou à blocs latents (LBM, aussi appelés SBM bipartites) sont des modèles de mélange particuliers dédiés au cas où les observations se présentent sous la forme d'un réseau.

Les modèles de Markov cachés (HMM en anglais, pour *Hidden Markov Model*) sont souvent utilisés pour analyser des données recueillies au cours du temps (par exemple, la trajectoire d'un animal observée en une série de dates) et font intervenir un processus sous-jacent (dans l'exemple, l'activité d'un animal : sommeil, déplacement, chasse) qui influe sur les observations (sa position ou sa trajectoire). Dans cet exemple-cas, les variables latentes sont discrètes et représentent l'activité de l'animal à chaque instant. Il existe également des modèles où le processus caché est lui-même continu.

Les modèles linéaires (généralisés) mixtes constituent aujourd'hui un outil de base en écologie pour décrire les effets d'un ensemble de conditions (environnementales, climatiques, etc.) sur une population ou une communauté. Ces modèles font intervenir des effets aléatoires qui sont, de fait, des variables latentes dans le but de prendre en compte soit une dispersion plus élevée qu'attendue, soit une dépendance entre les

observations. Le plus souvent, ces variables latentes sont continues et de nature plutôt instrumentale.

Les modèles de distributions jointes d'espèces (JSDM) sont une version multidimensionnelle des modèles linéaires généralisés visant à décrire la composition d'une communauté en fonction non seulement de variables environnementales mais aussi des interactions entre les espèces qui la composent. De nombreux JSDM ont recours à une variable latente multidimensionnelle (par exemple gaussienne) dont la structure de dépendance est censée décrire les interactions interspécifiques.

La modélisation en écologie vise fréquemment à décrire les effets de conditions expérimentales ou de variables environnementales sur la réponse ou le comportement d'une ou plusieurs espèces. De telles variables explicatives sont souvent appelées covariables. Ces effets sont typiquement pris en compte au moyen d'un terme de régression, comme par exemple dans les modèles linéaires généralisés. Il est généralement possible d'inclure un tel terme de régression dans les modèles à variables latentes qui prévoient alors que la distribution de la variable réponse étudiée dépende à la fois des covariables observées et de variables latentes (qui ne le sont pas).

I.3.2. *Estimation*

De nombreuses méthodes ont été proposées pour estimer les paramètres d'un modèle à variables latentes. Dans le cadre fréquentiste, la méthode la plus ancienne et la plus populaire pour obtenir les estimateurs du maximum de vraisemblance est sans doute l'algorithme Espérance-Maximisation (EM) qui tire parti du fait que, pour bon nombre de ces modèles, l'estimation des paramètres serait relativement facile si les variables latentes étaient observées. L'algorithme EM alterne l'étape E qui vise à calculer toutes les quantités impliquant les variables latentes nécessaires à la mise à jour des estimations des paramètres lors de l'étape M. L'étape E repose principalement sur la détermination de la distribution conditionnelle des variables latentes sachant les données observées. Ce calcul peut être immédiat (comme pour les modèles de mélange ou certains modèles mixtes), ou possible mais coûteux (comme pour les modèles de Markov cachés), mais il peut aussi s'avérer impossible pour des raisons combinatoires ou formelles.

Le problème de l'estimation se pose encore plus crûment en inférence bayésienne puisqu'il s'agit d'établir non seulement la distribution conditionnelle des variables latentes, mais également celle des paramètres. Là encore, sauf dans quelques cas très particuliers, la détermination de manière exacte de cette loi conditionnelle jointe (variables latentes et paramètres) est le plus souvent impossible.

Deux grandes familles de méthodes se dégagent pour traiter de l'inférence dans les modèles dans lesquels cette loi conditionnelle n'est pas calculable : les méthodes

d'échantillonnage et les méthodes d'approximation. Les méthodes de la première famille visent à obtenir un échantillon issu de cette loi inaccessible afin d'en déduire des estimations précises de toutes les quantités d'intérêt. Cette famille comprend notamment les algorithmes de Monte-Carlo, de Monte-Carlo par chaîne de Markov (MCMC), de Monte-Carlo séquentielles (SMC). Ces algorithmes sont, par nature, aléatoires. Ils sont particulièrement populaires en inférence bayésienne. Les méthodes de la seconde famille visent à déterminer une approximation de la loi conditionnelle des variables latentes (et des paramètres dans le cas bayésien) sachant les observations. Ces méthodes contiennent notamment les méthodes dites variationnelles et leurs extensions. Elles se distinguent entre elles par la mesure de la proximité entre la loi approchée et la vraie loi conditionnelle et par la famille de distributions au sein de laquelle l'approximation est recherchée.

I.4. Positionnement de l'ouvrage et grille de lecture

Cet ouvrage apporte un panorama des travaux récents sur la modélisation statistique et l'estimation dans les modèles à variables latentes pour l'écologie. Les différents chapitres sont des illustrations des grands principes qui viennent d'être décrits. Ils présentent des méthodes statistiques reposant sur des modèles et des algorithmes classiques dans certains cas, ou sur des développements issus de travaux de recherche récents dans d'autres. Chaque chapitre aborde une question écologique particulière, présente une manière de la traiter par la modélisation et illustre la démarche sur un ou plusieurs cas d'étude. De plus, le code R est mis à disposition[1], de manière à permettre à chacun de s'approprier les outils, et de les mettre en œuvre sur ses propres données.

Les questions associées aux cas d'étude sont principalement des questions de compréhension, de description des systèmes étudiés. Les questions de prédiction, même si elles sont abordées dans quelques chapitres, ne sont pas au cœur de l'ouvrage. Le cas des données manquantes (c'est-à-dire des valeurs non observées dans un échantillon) n'est pas abordé non plus. Enfin, cet ouvrage ne prétend pas être un ouvrage de synthèse sur les modèles à variable latentes et les méthodes et algorithmes d'inférence dédiés. Chaque chapitre aborde la question de l'inférence du modèle étudié, mais renvoie aux références appropriées pour les détails techniques.

Cet ouvrage n'est pas à lire de manière linéaire. Chacun ira piocher dans les chapitres en fonction de ses attentes d'écologue modélisateur ou de statisticien qui s'intéresse à l'écologie. Ainsi, les chapitres sont organisés par échelle écologique croissante, de l'individu aux écosystèmes, ce qui offre une première grille de lecture. Une

1. https://oliviergimenez.github.io/code_livre_variables_cachees/.

autre entrée peut se faire par la nature de la variable cachée que l'on cherche à modéliser. Enfin, l'entrée peut se faire par le modèle statistique : un même modèle peut être mobilisé dans plusieurs chapitres, pour des questions à des échelles différentes et avec différentes méthodes d'estimation mises en œuvre. Le tableau I.1 présente une vue synthétique du contenu des chapitres et permettra au lecteur de cibler ceux qui l'intéressent.

Chapitre	Échelle	Modèle	Variable latente	
			Nature	Domaine
1	Individu	Chaîne de Markov cachée	Cachée : position Fictive : comportement	$\mathbb{R}$ $\{1, \ldots, K\}$
2	Individu	Chaîne de Markov cachée	Fictive : signaux proximaux Cachée : acquisition et allocation de ressource	$\{1, \ldots, K\}$ $\{1, \ldots, K\}$
3	Population	Chaîne de Markov cachée	Cachée : dynamique de la population	$\{1, \ldots, K\}$
4	Population	EDO bruitée	Cachée : taille de la population	$\mathbb{N}^+$
5	Métapopulation	Chaîne de Markov cachée spatialisée	Cachée : classe de l'état dormant	$\{1, \ldots, K\}$
6	Communauté	Mélange (SBM et SBM bipartites)	Fictive : groupes d'espèces ayant la même structure d'interaction	$\{1, \ldots, K\}$
7	Communauté	Régression (JSDM sur présence-absence)	Fictive : corrélations entre espèces	$\mathbb{R}$
8	Communauté	Régression (JSDM sur comptage)	Fictive : corrélations entre espèces	$\mathbb{R}$
9	Communauté	Régression (JSDM sur comptage)	Instrumentale : composantes résumant les covariables	$\mathbb{R}$
10	Socio-écosystème	Régression (MES)	Cachée : composantes du système non directement mesurables et en interaction	$\mathbb{R}$

Tableau I.1. *Vue synthétique des chapitres*

I.5. Ouverture

Les exemples cités jusqu'ici et ceux développés dans les chapitres suivants montrent la très grande flexibilité offerte par les modèles à variables latentes. Ils

forment en effet un cadre très riche permettant, par l'inclusion de couche latentes plus ou moins nombreuses, de décrire des structures de dépendances parfois très complexes et/ou de s'approcher d'une description mécanistique des phénomènes.

Cette richesse ne doit cependant pas faire oublier que la sophistication du modèle s'accompagne presque toujours d'une complexité accrue de son inférence. Quelle que soit l'approche statistique adoptée (par exemple fréquentiste ou bayésienne), il serait faux de croire que « l'inférence suivra ».

Il n'existe pas à ce jour (et certainement pour longtemps encore) de technique complètement générique s'adaptant à tous les modèles. Même le recours à des algorithmes bien établis (EM, MCMC, etc.) nécessite souvent une bonne connaissance des principes qui les guident afin de bien contrôler leur comportement, voire des développements spécifiques. Les chapitres de cet ouvrage en sont autant d'exemples.

Nous conclurons cette introduction par l'évocation de deux pistes de recherche en écologie qui s'appuient sur la modélisation statistique de variables cachées, non abordées dans l'ouvrage et qui nous semblent prometteuses : l'intégration (ou combinaison) de données issues de plusieurs sources, et l'utilisation des données de sciences participatives.

L'intégration de données issues de plusieurs sources a fait l'objet de plusieurs travaux récemment en écologie (Isaac *et al*. 2020 ; Miller *et al*. 2019). Ces travaux sont motivés par l'amélioration systématique de la précision des paramètres estimés, la possible réduction des tailles d'échantillon ainsi que la perspective de pouvoir identifier des paramètres autrement non estimables. L'intégration de données repose en général sur une approche de modélisation hiérarchique dans laquelle la variable cachée est commune aux différentes sources qui servent à informer son estimation.

Les données de sciences participatives ont fait l'objet ces dernières années d'une attention croissante dans la littérature (Dickinson *et al*. 2012 ; McKinley *et al*. 2017). Cela s'explique par la prévalence croissance de ces données, ainsi que par la possibilité d'amasser de l'information à de larges échelles spatiales et temporelles. Ces données constituent un champ de recherche riche pour l'écologie statistique car elles présentent plusieurs défis comme les biais spatiaux d'échantillonnage ou encore les variations dans l'expertise des participants. Là encore, faire la distinction entre les processus écologiques d'intérêt représentés par les variables cachées et les processus d'observation associés permet de mieux répondre à la question écologique.

Nous vous souhaitons une bonne lecture.

I.6. Bibliographie

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T., Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, 10(6), 291–297.

Isaac, N.J., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., O'Hara, R.B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, 35(1), 56–67.

McKinley, D.C., Miller-Rushing, A.J., Ballard, H.L., Bonney, R., Brown, H., Cook-Patton, S.C., Evans, D.M., French, R.A., Parrish, J.K., Phillips, T.B., Ryan, S.F., Shanley, L.A., Shirk, J.L., Stepenuck, K.F., Weltzin, J.F., Wiggins, A., Boyle, O.D., Briggs, R.D., Chapin, S.F., Hewitt, D.A., Preuss, P.W., Soukup, M.A. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, 208, 15–28.

Miller, D.A.W., Pacifici, K., Sanderlin, J.S., Reich, B.J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1), 22–37.

Cette bibliographie est identique à celle de l'ouvrage correspondant en anglais publié par ISTE.