Les POMDP: une solution pour modéliser des problèmes de gestion adaptative en biologie de la conservation

Iadine Chadès, Josie Carwardine, Tara Martin, Samuel Nicol, Olivier Buffet

Les POMDP: une solution pour modéliser des problèmes de gestion adaptative en biologie de la conservation

Iadine Chadès[1], Josie Carwardine[1], Tara G. Martin[1], Samuel Nicol[2], Olivier Buffet[3]

[1] CSIRO Ecosystem Sciences
{iadine.chades, josie.carwardine, tara.martin}@csiro.au
[2] University of Alaska Fairbanks
snicol@alaska.edu

[3] INRIA / Nancy Université
olivier.buffet@loria.fr

Résumé :
En biologie de la conservation, la gestion adaptative est un processus itératif d'amélioration de la gestion par la réduction de l'incertitude à travers une surveillance. La gestion adaptative est l'outil principal pour la conservation d'espèces menacées par les changements planétaires, toutefois les problèmes de gestion adaptative souffrent d'un ensemble pauvre de méthodes de résolution. L'approche courante employée pour résoudre un problème de gestion adaptative est de faire l'hypothèse que l'état du système est connu et que sa dynamique est dans un ensemble de modèles pré-définis. La méthode de résolution utilisée n'est pas satisfaisante parce qu'elle emploie l'algorithme d'itération sur la valeur sur un belief MDP discrétisé qui restreint l'étude à de très petits problèmes. Nous montrons comment dépasser cette limitation en modélisant un problème de gestion adaptative par un type particulier de processus de décision markovien partiellement observable (POMDP) appelé MDP à observabilité mixte (MOMDP). Nous montrons comment simplifier la fonction de valeur, l'opérateur de mise à jour de la fonction de valeur et le calcul de mise à jour de l'état de croyance. Ceci ouvre la voie à des améliorations des algorithmes de résolution des POMDP. Nous illustrons l'utilisation de notre MOMDP "adaptatif" à la gestion d'une population de pinsons diamants de Gould, une espèce d'oiseaux endémique de l'Australie du nord. Notre approche de modélisation simple est une grande avancée pour la résolution de problèmes de gestion adaptative pour la conservation en utilisant des méthodes efficaces pour les POMDP.

Introduction

En biologie de la conservation, la gestion adaptative – ou "apprendre en faisant" – est un processus itératif de réduction progressive de l'incertitude à l'aide de la surveillance. Inventé par Walters & Hilborn (1978), la gestion adaptative a gagné en notoriété en tant qu'approche pour gérer des écosystèmes afin de préserver leur biodiversité. Les éléments clefs de la gestion adaptative incluent une définition claire des objectifs, une spécification de modèles alternatifes pour atteindre ces objectifs, l'implémentation d'au moins deux modèles dans un cadre d'expérimentation comparative, la surveillance pour évaluer les mérites relatifs et limitations des modèles alternatifs, et la modification itérative de la gestion pour déterminer le modèle réel, s'il y en a un (Keith *et al.*, In press). Malgré ses vertues, il y a peu d'exemples d'applications pratiques de la gestion adaptative. Plusieurs facteurs ont été proposés pour expliquer les difficultés d'implémentation courantes dans les programmes de gestion adaptative. Parmi eux on compte l'inefficacité des méthodes utilisées pour résoudre les problèmes de gestion adaptative et l'incapacité à manipuler expérimentalement les populations menacées.

A ce jour, les problèmes de gestion adaptative ont été abordés en utilisant les méthodes pour belief MDP discrétisés (Williams, 2009). L'espace de croyance est discrétisé en utilisant p sous-intervalles pour chaque variable de croyance. La règle de mise à jour ne garantit pas que la croyance mise à

jour tombe sur l'un des points de cette grille, ce qui nécessite l'utilisation d'une règle d'interpolation pour définir les probabilités de transition pour les états de croyance. Cette technique a aussi été étudiée pour réduire le coût computationnel de l'algorithme value iteration exact pour les POMDP, en utilisant soit une grille fixe (Lovejoy, 1991; Bonet, 2002) soit une grille variable (Brafman, 1997; Zhou & Hansen, 2001). Les méthodes à base de grilles diffèrent principalement par la méthode de sélection des points et par la forme de la fonction d'interpolation. En général, les grilles régulières ne passent pas bien à l'échelle dans les problèmes de grande dimension et les grilles non régulières souffrent de lourds calculs d'interpolation.

Dans cet article, nous proposons une approche transparente et formelle pour représenter un problème de gestion adaptative comme un cas particulier de POMDP. Nous montrons pour la première fois comment modéliser et résoudre un problème de gestion adaptative comme un MDP à observabilité mixte (MOMDP). Notre approche bénéficie des développements récents dans le domaine de la robotique et de la prise de décision dans l'incertain (Ong *et al.*, 2010; Araya-López *et al.*, 2010). Notre objectif est de faire connaître à la communauté de l'IA un problème d'optimisation stimulant mais résoluble dans le domaine de la biologie de la conservation. Notre méthode est particulièrement pertinente dans le cas d'espèces menacées pour lesquelles il peut ne pas être possible de conduire des expérimentations répétées pour apprendre leur vrai modèle. Nous croyons que des développements méthodologiques supplémentaires amélioreront l'intérêt de la gestion adaptative à travers la capacité à clairement articuler le problème de la gestion adaptative.

Cas d'emploi

Nous illustrons notre méthode sur la gestion d'une population d'une espèce d'oiseaux menacée, le diamant de Gould. Les menaces les plus répandues sur les populations sauvages de diamants de Gould sont la perte et la dégradation d'habitat causés par des feux et des régimes de pâturage inappropriés, et par l'introduction de prédateurs tels que des chats sauvages. La réponse de la population aux différentes actions de gestion est incertaine. Chacun de nos quatre experts a fourni un modèle possible sous la forme de distributions de probabilitéss décrivant comment la population pourrait répondre à quatre actions de gestion de menaces. Notre objectif est d'implémenter l'action de gestion qui mènera le plus vraisemblablement vers une forte probabilité de persistence de la population de diamants de Gould. Une stratégie adaptative optimale fournira la meilleure décision pour déterminer au bout d'un certain temps quel modèle (entre 1 et 4) est plus vraisemblablement le modèle réel, et ainsi quelle action est optimale.

POMDP

Les POMDP sont un modèle commode pour la résolution de problèmes de prise de décision séquentielle optimale quand le décideur n'a pas d'information complète à propos de l'état courant du système.

Formellement, un POMDP discret avec horizon infini est spécifié par un uplet $\langle S, A, O, T, Z, r, \gamma \rangle$, où :

- S est l'ensemble des états s qui peuvent être partiellement observés ou détectés imparfaitement par le gestionnaire ;
- A est l'ensemble des actions (ou décisions) a parmi lesquelles le gestionnaire a besoin de choisir à chaque pas de temps ;
- T est une fonction de transition probabiliste décrivant la dynamique stochastique du système ; $T(s, a, s')$ représente la probabilité d'être dans l'état s' au temps $t + 1$ étant donné (s, a) à t : $T(s, a, s') = p(s_{t+1} = s' | s_t = s, a_t = a)$;
- O est l'ensemble des observations o que le gestionnaire peut percevoir ;
- Z est la fonction d'observation, où $Z(a, s', o') = p(o_{t+1} = o' | a_t = a, s_{t+1} = s')$ représente la probabilité conditionnelle que le gestionnaire observe o' sachant que l'action a a mené à l'état s' ;
- $r : S \times A \to \Re$ est la fonction de récompense identifiant les bénéfices et coûts d'être dans un état particulier et d'effectuer une action ;
- $\gamma \in [0, 1)$ est un facteur d'atténuation.

La décision optimale au temps t peut dépendre de l'historique complet des actions et observations passées. Parce qu'il n'est pas pratique d'utiliser l'historique des trajectoires observation-action pour calculer ou représenter une solution optimale, des états de croyance – c'est-à-dire des distributions de probabilités sur les états – sont utilisées pour résumer les informations détenues et surmonter les difficultés d'une détection imparfaite. Un POMDP peut être transformé en un processus de décision markovien totalement observable défini sur l'espace (continu) des états de croyance. Dans notre cas, résoudre un POMDP signifie trouver une stratégie $\pi : B \rightarrow A$ associant à l'état de croyance courant ($b \in B$) une allocation des ressources. Une stratégie optimale maximise l'espérance de la somme des récompenses atténuées sur un horizon temporel infini $E[\sum \gamma^t R(b_t, a_t)]$ où b_t et a_t désignent l'état de croyance et l'action au temps t, et $R(b,a) = \sum_s b(s)r(s,a)$. Pour un état de croyance b et une politique donnée π cette somme espérée est aussi appelée la fonction de valeur $V_\pi(b)$. Une fonction de valeur nous permet de classer les stratégies en affectant une valeur réelle à chaque croyance b. Une stratégie optimale π^* est une stratégie telle que $\forall b \in B \ \forall \ \pi \ V_{\pi^*}(b) \geq V_\pi(b)$. Plusieurs stratégies peuvent être optimales et partager la même fonction de valeur optimale V^*, laquelle peut être calculée en utilisant l'opérateur de programmation dynamique pour un POMDP représenté comme un belief MDP (Bellman, 1957) :

$$\forall b \in B, V^*(b) = \max_{a \in A} \left[\sum_{s \in S} r(s,a)b(s) + \gamma \sum_{o'} p(o'|b,a)V^*(b^{ao'}) \right],$$

où $b^{ao'}$ est la croyance mise à jour étant donné que l'action a a été effectuée et que o' est observé. Cette fonction peut être calculée récursivement en utilisant le principe d'optimalité de Bellman (Bellman, 1957) :

$$V_{n+1}(b) = \max_{a \in A} \left[\sum_{s \in S} r(s,a)b(s) + \gamma \sum_{o'} p(o'|b,a)V_n(b^{ao'}) \right]. \quad (1)$$

$b^{ao'}$ est mis à jour en utilisant la règle de Bayes :

$$b^{ao'}(s') = \frac{p(o'|a,s')}{p(o'|b,a)} \sum_{s \in S} p(s'|s,a)b(s). \quad (2)$$

Sondik (1971) a montré que la fonction de valeur pour horizon temporel fini est linéaire par morceaux et convexe (PWLC pour "PieceWise Linear and Convex") et que la fonction de valeur pour horizon temporel infini peut être approchée avec une précision arbitraire par une fonction PWLC. La convexité implique que la valeur d'une croyance proche de l'un des coins du simplex de croyance B sera élevé. Une représentation alternative de V consiste à utiliser des vecteurs :

$$V(b) = \max_{\alpha \in \Gamma} \alpha \cdot b, \quad (3)$$

où Γ est un ensemble fini de vecteurs appelés α-vecteurs, b est la croyance représentée comme un vecteur de dimension finie, et $\alpha \cdot b$ dénote le produit scalaire entre un α-vecteur et b. Le gradient de la fonction de valeur à b est donnée par le vecteur $\alpha_b = \arg\max_{\alpha \in \Gamma} b \cdot \alpha$. La politique peut être exécutée en évaluant (3) à b pour trouver le meilleur α-vector : $\pi(b) = a(\alpha_b)$. Les méthodes exactes comme Incremental Pruning (IP) reposent sur l'élagage régulier des hyperplans dominés pour réduire leur nombre (Cassandra, 1998).

Alors que divers algorithmes de la littérature en recherche opérationnelle et en intelligence artificielle ont été développés ont été développés ces dernières années, la complexité computationnelle des algorithmes exacts reste insoluble pour la plupart des problèmes : les POMDP à horizon fini sont PSPACE-complets (Papadimitriou & Tsitsiklis, 1987) et les POMDP à horizon infini sont indécidables (Madani *et al.*, 2003).

Ces quelques dernières années, des méthodes approchées ont été développées avec succès pour résoudre des POMDP de grande taille. Parmi eux, les approches à base de points approchent la fonction de valeur en ne la mettant à jour que pour certains états de croyance sélectionnés (Pineau *et al.*, 2003; Spaan & Vlassis, 2005). Les méthodes à base de points typiques échantillonnent les

états de croyance en simulant des interactions dans l'environnement puis mettent à jour la fonction de valeur et son gradient sur une sélection des états de croyance ainsi échantillonnés.

Une autre amélioration notable est l'usage de POMDPs factorisés et de méthodes de résolution associées (Boutilier & Poole, 1996). Un POMDP factorisé est un POMDP dans lequel les relations d'indépendence entre les variables du systèmes sont représentées explicitement.

Les MOMDP pour la gestion adaptative

Dans cette section nous montrons comment modéliser des problèmes de gestion adaptative en employant un modèle de POMDP factorisé spécifique.

MOMDP

Un modèle MOMDP (Ong *et al.*, 2010) est spécifié comme un uplet $\langle X, Y, A, O, T_x, T_y, Z, r, \gamma \rangle$ où :

- $S = X \times Y$ est l'ensemble factorisé des états du système avec X la variable aléatoire représentant les composantes complètement observables et Y la variable aléatoire représentant les composantes partiellement observables. Dans notre cadre, Y représente le modèle inconnu. Une paire (x, y) spécifie l'état du système naturel complet. Cette représentation factorisée permet une représentation plus structurée et compacte des fonctions de transition, d'observation et de récompense, ainsi que des calculs plus efficaces ;
- A est l'ensemble fini des actions ;
- $T_y(y, a, y') = p(y'|y, a)$ donne la probabilité que la valeur de la variable d'état partiellement observable change de y à y' si l'action a est effectuée ; $T_x(x, y, a, x', y') = p(x'|x, y, a, y')$ donne la probabilité que la variable d'état totalement observable prenne la valeur x' au temps $t + 1$ si l'action a est effectuée dans l'état (x, y) au temps t et a déjà mené à y' ;
- $O = O_x \times O_y$ est l'ensemble des observations avex $O_x = X$ la composante complètement observable, et O_y l'ensemble des observations des variables cachées ;
- Z est la fonction d'observation – aussi appelée fonction de probabilité de détection en écologie – avec $Z(a, x', y', o'_x, o'_y) = p(o'_x, o'_y|a, x', y')$ la probabilité d'observer o'_x, o'_y si l'action a a été exécutée, menant à l'état du système (x', y'). Dans un MOMDP la variable X est parfaitement observable (et donc la valeur x' connue) de sorte que nous avons $p(o'_x|a, x', y', o'_y) = 1$ si $o'_x = x'$, et 0 sinon ;
- $r : S \times A \to \Re$ est la fonction de récompense usuelle définie sur la variable d'état S et l'ensemble d'actions A ;
- $\gamma \in [0, 1)$ est un facteur d'atténuation.

L'espace de croyance B ne concerne ici que notre croyance sur y uniquement puisque la variable d'état X est complètement observable. Toute croyance $b \in B$ sur l'état complet du système $s = (x, y)$ est représenté par (x, b_y). Dans les algorithmes reposant sur des approximations PWLC, prendre en compte les variables d'état visibles permet des accélérations notables en raisonnat sur plusieurs espaces de croyance de faible dimensionalité au lieu de l'espace original de haute-dimensionalité, comme démontré avec MO-SARSOP (Ong *et al.*, 2010) – basé sur SARSOP, un résolveur à base de points de l'état de l'art – et MO-IP (Araya-López *et al.*, 2010) – basé sur Incremental Pruning.

MOMDP appliqués à la gestion adaptative

Notre problème de gestion d'adaptative suppose que les gestionnaires peuvent parfaitement observer l'état du système naturel étudié mais sont incertains en ce qui concerne la dynamique du système. Nous listons les hypothèses qui font de l'application des MOMDP aux problèmes de gestion adaptative un défi potentiellement plus facile. Les hypothèses communes qui suivent peuvent être incorporées dans les MOMDP pour simplifier la solution :

1. le modèle réel des dynamiques du système y_r est un élément d'un ensemble fini de modèles prédéfinis Y ;

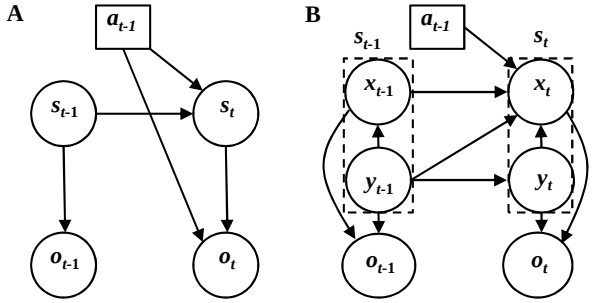

FIGURE 1 – Le POMDP standard (A) comparé avec notre MOMDP pour gestion adaptative (B).

2. l'ensemble fini des actions A affecte la variable complètement observable X et n'affecte pas Y : $p(y'|y,a) = p(y'|y)$;
3. le modèle réel y_r – bien qu'inconnu – ne changera pas à travers le temps et donc T_y est la matrice identité, c'est-à-dire que $p(y'|y) = 1$ si $y = y'$ et 0 sinon ;
4. la variable cachée Y ne peut être observée et la fonction d'observation Z est seulement définie sur la variable complètement observable X, c'est-à-dire que $O = O_x$ et $p(o'_x|a,x') = 1$ si $o'_x = x'$ et 0 sinon.

La figure 1 illustre la comparaison du modèle POMDP standard avec le modèle MOMDP pour notre application à la gestion adaptative. Une variante des problèmes de gestion adaptative est le cas dans lequel on fait l'hypothèse que le modèle réel y_r peut changer avec le temps. En d'autres termes T_y n'est plus la matrice identité et permet de transiter entre états non observés y : la probabilité de transition d'un modèle y à un modèle $y' \neq y$ peut être non nulle. En écologie, cela pourrait être le cas quand on a affaire avec avec des systèmes non stationaires ou périodiques tels que le climat, les précipitations ou les températures. Une autre variante des problèmes de gestion adaptative est de permettre une observabilité partielle de l'espace d'état X. Dans ce cas le problème de gestion adaptative sera un POMDP. On discutera de ces hypothèses dans la section discussion.

Méthodes de résolution

En nous appuyant sur (Ong *et al.*, 2010) et (Araya-López *et al.*, 2010), nous motrons comment simplifier le calcul de la fonction de valeur (1) et la mise à jour de la croyance (2) pour des algorithmes de résolution de POMDP existants. Dans (Ong *et al.*, 2010), les auteurs ont montré que $V(b) = V(x, b_y)$ pour tout MOMDP, et l'équation (1) peut être ré-écrite :

$$V_{n+1}(x,b_y) = \max_{a\in A}[\sum_{y\in Y} r(x,a)b_y(y) + \gamma \sum_{y,x',y',o'} b_y(y)p(x'|x,y,a,y')p(y'|y,a)p(o'|a,x',y')V_n(x',b_y^{ao'})]. \tag{4}$$

En tenant compte des hypothèse de stationarité (2-3) sur Y, nous pouvons simplifier (4) :

$$V_{n+1}(x,b_y) = \max_{a\in A}[\sum_{y\in Y} r(x,a)b_y(y) + \gamma \sum_{y,x',o'} b_y(y)p(x'|x,y,a)p(o'|x',y)V_n(x',b_y^{ao'})] \tag{5}$$

En tenant compte de $p(o'|x',y') = 1$, $o' = x'$, l'hypothèse (4) mène à la simplification :

$$V_{n+1}(x,b_y) = \max_{a\in A}[\sum_{y\in Y} r(x,a)b_y(y) + \gamma \sum_{y,x'} b_y(y)p(x'|x,y,a)V_n(x',b_y^{a,o'=x'})], \tag{6}$$

qui peut être ré-écrite en substituant $V_n(x', b_y^{a,o'=x'}) = \max_{\alpha\in\Gamma_n} b_y^{a,o'=x'} \cdot \alpha$:

$$V_{n+1}(x,b_y) = \max_{a\in A}[\sum_{y\in Y} r(x,a)b_y(y) + \gamma \sum_{y,x'} b_y(y)p(x'|x,y,a) \max_{\alpha\in\Gamma_n} b_y^{a,o'=x'} \cdot \alpha] \tag{7}$$

Algorithme 1 : Mise à jour de l'α-vecteur par MO-SARSOP au nœud (x, b_y) de T_R

1 BACKUP$(T_R, \Gamma, (x, b_y))$
2 **forall** $a \in A, x' \in X, o' \in O$ **do**
3 $\quad \alpha_{a,x',o'} \leftarrow \arg\max_{\alpha \in \Gamma_y(x')}(\alpha \cdot \tau(x, b_y, a, x', o'))$
4 **forall** $a \in A, y \in Y$, **do**
5 $\quad \alpha_a(y) \leftarrow r(x, y, a) + \gamma \sum_{x',o',y'} T_x(x, y, a, x') T_y(x, y, a, x', y') \times Z(x', y', a, o') \alpha_{a,x',o'}(y')$
6 $\alpha' \leftarrow \arg\max_{a \in A}(\alpha_a \cdot b_y)$
7 Insérer α' dans $\Gamma_y(x)$

avec $b_y^{a,o'=x'}$ la croyance mise à jour étant donné a, o' et b_y simplifiée comme suit :

$$b_y^{a,o'=x'}(y) = \frac{p(x'|x, y, a) b_y(y)}{\sum_{y''} b_y(y'') p(x'|x, y'', a)}. \tag{8}$$

En substituant (8) dans $\max_{\alpha \in \Gamma_n} b_y^{a,o'=x'} \cdot \alpha$ on peut réarranger (7)

$$\begin{aligned} V_{n+1}(x, b_y) &= \max_{a \in A} [\sum_{y \in Y} r(x, a) b_y(y) + \gamma \sum_{y,x'} b_y(y) p(x'|x, y, a) \frac{\max_{\alpha \in \Gamma_n} \sum_{y'} p(x'|x, y', a) b_y(y') \alpha(y')}{\sum_{y''} b_y(y'') p(x'|x, y'', a)}] \\ &= \max_a [\sum_{y \in Y} r(x, a) b_y(y) + \gamma \sum_{x'} \max_{\alpha \in \Gamma_n} \sum_{y'} p(x'|x, y', a) b_y(y') \alpha(y')] \\ &= \max_a [b_y \cdot \alpha_0^a + \gamma \sum_{x'} \max_{\alpha \in \Gamma_n} b_y \cdot G_\alpha^{ax'}(x)], \end{aligned} \tag{9}$$

où $G_\alpha^{ax'}(x) = \sum_{y'} p(x'|x, y', a) \alpha(y')$ et la fonction de récompense $r(s, a)$ est représentée comme l'ensemble des $|A|$ vecteurs $\alpha_0^a = (\alpha_0^a(1), ..., \alpha_0^a(|S|))$, un par action a, $\alpha_0^a(s) = r(s, a)$.

En utilisant l'identité $\max_{y_j} x \cdot y_j = x \cdot \arg\max_{y_j} x \cdot y_j$ deux fois on obtient :

$$\begin{aligned} V_{n+1}(x, b_y) &= \max_a [b_y \cdot \alpha_0^a + \gamma b_y \cdot \sum_{x'} \arg\max_{\alpha \in \Gamma_n} b_y \cdot G_\alpha^{ax'}(x)] \\ &= b_y \cdot \arg\max_a b_y \cdot [\alpha_0^a + \gamma \sum_{x'} \arg\max_{\alpha \in \Gamma_n} b_y \cdot G_\alpha^{ax'}(x)]. \end{aligned}$$

Finalement on peut définir le vecteur backup(b_y) comme le vecteur dont le produit scalaire avec b_y fournit $V_{n+1}(b_y)$:

$$\begin{aligned} &V_{n+1}(x, b_y) = b_y \cdot \text{backup}(b_y), \text{ où} \\ &\text{backup}(b_y) = \arg\max_a b_y \cdot [\alpha_0^a + \gamma \sum_{x'} \arg\max_{\alpha \in \Gamma_n} b_y \cdot G_\alpha^{ax'}(x)]. \end{aligned} \tag{10}$$

En simplifiant le calcul de l'opérateur de mise à jour (10) et l'opération de mise à jour de la croyance (8) on a montré comment des algorithmes de résolution de POMDP à base de points basés sur une approximation PWLC peuvent être adaptés dans le cas de problèmes de gestion adaptative. Ces deux procédures sont communes à deux algorithmes exacts tels que Witness ou Incremental Pruning (Cassandra, 1998), et à la plupart des algorithmes à base de points, tels que PBVI (Pineau *et al.*, 2003), Perseus (Spaan & Vlassis, 2005), symbolic Perseus (Poupart, 2005) et SARSOP (Ong *et al.*, 2010).

L'algorithme 1 présente la procédure de mise à jour employée à chaque point de croyance par MO-SARSOP, et l'algorithme 2 montre comment – sous les hypothèses de gestion adaptatives – elle peut être simplifiée à deux niveaux : i) on n'a pas besoin de considérer l'ensemble d'observations O, et le calcul de la mise à jour de la croyance τ est simplifié (ligne 2) ; ii) la fonction d'observation Z et le modèle de dynamique T_y ne sont pas requis (ligne 5). La procédure d'échantillonnage profite aussi de la mise à jour simplifiée de la croyance (8).

Algorithme 2 : Mise-à-jour de l'α-vecteur par Adaptive MO-SARSOP au nœud (x, b_y) de T_R

BACKUP$(T_R, \Gamma, (x, b_y))$
forall $a \in A, x' \in X$ **do**
  $\alpha_{a,x'} \leftarrow \arg\max_{\alpha \in \Gamma_y(x')}(\alpha \cdot \tau(x, b_y, a, x'))$
forall $a \in A, y \in Y$, **do**
  $\alpha_a(y) \leftarrow r(x, a) + \gamma \sum_{x'} T_x(x, y, a, x')\alpha_{a,x'}(y)$
$\alpha' \leftarrow \arg\max_{a \in A}(\alpha_a \cdot b_y)$
Insérer α' dans $\Gamma_y(x)$

	Problème 1 $\|X\| = 2, \|Y\| = 4$	Problème 2 $\|X\| = 81, \|Y\| = 2$
IP	h= 5 $\|\alpha\| = 3753$ t=349,2s	h=4 $\|\alpha\| = 1181$ t=703,7s
MO-IP	h=5 $\|\alpha\| = 2052$ t=106,8s	h=4 $\|\alpha\| = 218$ t=0,59s
Grid bMDP$^+$	h=26 $\|\alpha\| = 4402$ t=1831s err=1,28	h=23 $\|\alpha\| = 426$ t=1849s err= 3,84
Symbolic Perseus*	$\|\alpha\| = 189$ $\bar{r} = 84,7$ t=59,2s	$\|\alpha\| = 28$ $\bar{r} = 73,78$ t=71,4s
SARSOP*$^+$	$\|\alpha\| = 5520$ $\bar{r} = 93,5$ t=59,3s err=0,19	$\|\alpha\| = 4483$ $\bar{r} = 75,49$ t=72,4s err=0,0016
MO- SARSOP*$^+$	$\|\alpha\| = 8179$ $\bar{r} = 94,1$ t=59,2s err=0,17	$\|\alpha\| = 3299$ $\bar{r} = 75,7$ t=12,23s err<0,001

TABLE 1 – Performance pour les problèmes du diamant de Gould. Les expérimentations ont été conduites sur un ordinateur Core 2 à 2,40 GHz Core 2 avec 3,45 Go de mémoire. (*) Récompense moyenne $\bar{r}$ calculée sur 500 simulations de 50 pas de temps. ($^+$) Erreur estimée (err) fournit par les logiciels respectifs.

Gestion d'un oiseau menacé

Nous appliquons notre modèle de gestion adaptative à la population de diamants de Gould menacéé dans le Kimberley en Australie. Notre objectif de gestion est de maximiser la vraisemblance d'une haute probabilité de persistance de cette population à moyen terme, dans une zone où nous pouvons mesurer l'état de la population mais sommes incertains en ce qui concerne la réponse de l'état aux actions de gestion. Nous avons demandé à quatre expertes d'évaluer la vraisemblance d'une haute (respectivement basse) probabilité de persistance sous quatre actions de gestion possibles. Suivant notre modèle MOMDP pour la gestion adaptative, nous définissons l'ensemble d'états $S = X \times Y$ où X représente la probabilité locale de persistance $X = \{$Low, High$\}$. Nous posons $Y = \{$Expert 1, Expert 2, Expert 3, Expert 4$\}$ l'ensemble des modèles réels possibles qui prédisent la dynamique de l'état de l'espèce que nous étudions. L'ensemble des actions $A = \{$DN,FG,C,N$\}$ représente les actions de gestion parmie lesquelles choisir à chaque pas de temps ; ne rien faire (DN), [1] améliorer la gestion du feu et des pâturages (FG), [2] contrôler les chats sauvages (C), et fournir des boîtes de nidification (N). [3] Nous faisons l'hypothèse que la même quantité de fonds sont dépensés pour chaque action. Les experts fournissent des probabilités de transition T_x pour chaque modèle et transition d'état. Finalement nous définissons la fonction de récompense de sorte que r(Low,DN) = 0, r(High,DN) = 20, r(Low,$\{$FG,C,N$\}$) = -5 et r(High,$\{$FG,C,N$\}$) = 15. Nous faisons l'hypothèse que, au début de notre programme de gestion, chaque expert a autant de chances d'avoir raison.

1. Do Nothing
2. Fire and Grazing
3. Nesting boxes

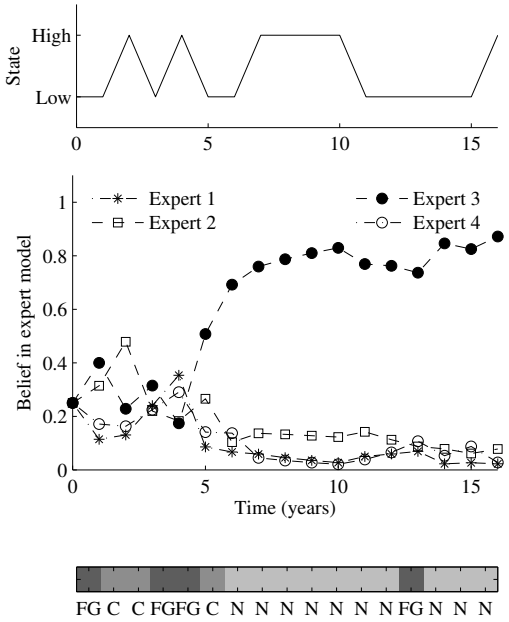

FIGURE 2 – Simulation de notre stratégie de gestion adaptative quand l'expert 3 détient le modèle réel. Les graphes représentent la dynamique de l'état (haut), la croyance en les prédictions de chaque expert à travers le temps (milieu), et l'action effectuée (bas).

Nous avons d'abord tenté de résoudre notre problème de gestion adaptative en utilisant IP.[4] Du fait d'un manque d'espace mémoire nous n'avons pas pu fournir une solution exacte pour un horizon temporel au-delà de 5 pas de temps avec 3753 α-vecteurs calculés (Problème 1, Table 1). Nous avons résolu le même problème avec MO-IP (Araya-López *et al.*, 2010) et atteint le 5ième horizon 3 fois plus vite qu'IP. Nous avons alors calculé des solutions approchées en utilisant les algorithmes grid-bMDP[4], Symbolic Perseus (Poupart, 2005), SARSOP and MO-SARSOP (Ong *et al.*, 2010). Nous avons interrompu Grid-bMDP après ∼30 minutes (horizon 26, 4402 α-vecteurs et une erreur estimée de 1,28). Symbolic Perseus a fournit une solution en moins d'une minute avec 189 α-vecteurs et une récompense moyenne de $\bar{r} = 84,7$. SARSOP et MO-SARSOP ont fournit de loin les meilleures performances en moins de 60s.[5] Les expérimentations menées avec une version de MO-IP implémentant les modifications spécifiques à la gestion adaptative n'ont pas montré d'accélération à l'horizon 5 (106,69s contre 106,80s pour le MO-IP original). Nous avons aussi lancé ces algorithmes sur une version plus complexe du problème du diamant de Gould où l'on distingue explicitement les interactions entre 4 espèces menant à un total de $|X| = 81$ combinaisons d'états, 2 experts et 4 actions (problem 2, Table 1). Sans surprise, les algorithmes à base de points nous permettent de résoudre ce que l'on pensait être un problème de gestion adaptative insoluble avec des gains en performances évidents quand on exploite l'observabilité mixte.

Les solutions approchées du problème 1 fournissent des conseils sur l'action de gestion qui devrait être effectuée en l'absence d'une connaissance précise du modèle réel. Les experts 1 et 4 fournissent des probabilités de transition similaires ; les deux pensent que la gestion appropriée des feux et pâturages (FG) générera une plus haute probabilité que la population atteigne et maintienne in "Haut" état de persistance. En conséquence, en l'absence d'information a priori sur ce qu'est le modèle réel, la première action à effectuer est FG. L'expert 2 pense que la gestion des chats sauvages (C) et la fourniture de nichoirs (N) ont plus d'avantages que la gestion des feux et pâturages. L'expert 3 pense que la perte de l'habitat et la compétition entre espèces est le facteur limitant de la persistance et attribue pour cela les plus hautes probabilités de persistance quand

4. En utilisant la boîte à outils pomdp-solver de Cassandra.

5. Notons que Symbolic Perseus emploie MATLAB alors que tous les autres algorithmes sont en C/C++. Dans le cas de Symbolic Perseus, SARSOP et MO-SARSOP, nous avons utilisé la fonction de simulation fournie par chaque logiciel.

l'action "fournir des boîtes de nidification" (N) est mise en œuvre. Notons que l'expert 3 évalue la gestion des chats sauvages et des feux et pâturages comme de "bonnes" actions. Nous avons évalué la qualité de notre stratégie adaptative par simulation. Quand nous faisons l'hypothèse que les chats sont la principale menace et que l'expert 2 détient le modèle réel, notre stratégie adaptative trouve rapidement la meilleure action à effectuer. Pourtant quand on fait l'hypothèse que feux et pâturages sont les principales menaces, notre stratégie adaptative fournit la meilleure action à effectuer mais a du mal à identifier le modèle réel, parce que les experts 1 et 4 soutiennent de manière similaire l'action FG combinée. Quand on fait l'hypothèse que l'expert 3 détient le modèle réel (voir figure 2), les simulations amènent d'abord à croire que la gestion des chats est la plus avantageuses et que l'expert 2 détient le modèle réel. De nouveau cette ambiguité est due au haut taux de succès donné par l'expert 3 quand les chats sauvages sont gérés. Après plusieurs pas de temps, notre stratégie adaptative favorise l'expert 3.

Discussion

Les problèmes de gestion adaptative tels que résolus actuellement dans la littérature écologique souffrent du manque de méthodes de résolution efficaces. Nous montrons ici et expliquons comment modéliser un problème de gestion adaptative classique comme un POMDP avec observabilité mixte. Les POMDP sont un modèle puissant qui ne requiert pas les hypothèses restrictives de la littérature traditionnelle en gestion adaptative. Premièrement, il n'y a pas besoin de faire l'hypothèse d'un modèle statique (hypothèse 3). Si nous permettons au modèle réel de changer avec le temps nous pouvons aborder des systèmes qui sont influencés par le changement climatique ou des fluctuations saisonnières. Deuxièmement, il n'y a pas besoin de faire l'hypothèse d'une détection parfaite de l'état (X) (hypothèse 4) et les probabilités de détection peuvent être incluses. L'hypothèse que le modèle réel est parmi un ensemble prédéfini de modèles (hypothèse 1) reste un défi irrésolu. A notre connaissance cette hypothèse ne peut être abordée avec les solutions actuelles pour les POMDP mais l'apprentissage par renforcement bayésien (Poupart *et al.*, 2006) pourrait être une alternative pertinente. Dans les premiers temps de la gestion adaptives (Walters & Hilborn, 1978), nous n'étions pas capables de résoudre des POMDP efficacement, mais de récents algorithmes à base de points font des POMDP le couteau suisse des biologistes de la conservation (Chadès *et al.*, 2008; MacKenzie, 2009).

Références

Araya-López M., Thomas V., Buffet O. & Charpillet F. (2010). A closer look at MOMDPs. In *Proc. of the 22nd Int. Conf. on Tools with Artificial Intelligence*.

Bellman R. E. (1957). *Dynamic Programming*. Princeton, N.J. : Princeton University Press.

Bonet B. (2002). An e-optimal grid-based algorithm for partially obserable Markov decision processes. In *Proc. of the 19th Int. Conf. on Machine Learning (ICML-02)*.

Boutilier C. & Poole D. (1996). Computing optimal policies for partially observable decision processes using compact representations. In *Proc. of the Nat. Conf. on Artificial Intelligence*.

Brafman R. (1997). A heuristic variable grid solution method for POMDPs. In *Proc. of the Nat. Conf. on Artificial Intelligence*.

Cassandra A. R. (1998). *Exact and Approximate Algorithms for Partially Observable Markov Decision Processes*. PhD thesis, Brown University, Dept. of Computer Science.

Chadès I., McDonald-Madden E., McCarthy M. A., Wintle B., Linkie M. & Possingham H. P. (2008). When to stop managing or surveying cryptic threatened species. *PNAS*, **105**, 13936–13940.

Keith D. A., Martin T. G., McDonald-Madden E. & Walters C. (In press). Uncertainty and adaptive management for biodiversity conservation. *Biological Conservation*.

Lovejoy W. (1991). Computationally feasible bounds for partially observed Markov decision processes. *Operations research*, **39**(1), 162–175.

MacKenzie D. (2009). Getting the biggest bang for our conservation buck. *Trends in Ecology & Evolution*, **24**(4), 175–177.

Madani O., Hanks S. & Condon A. (2003). On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, **147**(1-2), 5–34.

Ong S. C. W., Png S. W., Hsu D. & Lee W. S. (2010). Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research*, **29**(8), 1053–1068.

Papadimitriou C. H. & Tsitsiklis J. N. (1987). The complexity of Markov decision processes. *Journal of Mathematics of Operations Research*, **12**(3), 441–450.

Pineau J., Gordon G. & Thrun S. (2003). Point-based value iteration : An anytime algorithm for POMDPs. In *Proc. of the Int. Joint Conf. on Artificial Intelligence*.

Poupart P. (2005). *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*. PhD thesis, University of Toronto.

Poupart P., Vlassis N., Hoey J. & Regan K. (2006). An analytic solution to discrete Bayesian reinforcement learning. In *Proc. of the 23rd Int. Conf. on Machine Learning*.

Sondik E. (1971). *The Optimal Control of Partially Observable Markov Decision Processes*. PhD thesis, Stanford University, California.

Spaan M. & Vlassis N. (2005). Perseus : Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, **24**, 195–220.

Walters C. J. & Hilborn R. (1978). Ecological optimization and adaptive management. *Annual Review of Ecology and Systematics*, **9**, pp. 157–188.

Williams B. (2009). Markov decision processes in natural resources management : Observability and uncertainty. *Ecological Modelling*, **220**(6), 830–840.

Zhou R. & Hansen E. (2001). An improved grid-based approximation algorithm for POMDPs. In *Proc. of the 17th Int. Joint Conf. on Artificial Intelligence*.