Apprentissage par renforcement pour l'optimisation de la conduite de culture du colza.

Ronan Trépos[1], Stéphane Lemarié[2], Hélène Raynal[1], Muriel Valantin-Morison[3], Stéphane Couture[1], Frédérick Garcia[1]

[1] INRA unité MIAT - Toulouse - Ronan.Trepos@toulouse.inra.fr, Helene.Raynal@toulouse.inra.fr, scouture@toulouse.inra.fr, fgarcia@toulouse.inra.fr
chemin de Borde-Rouge - 31326 Castanet-Tolosan Cedex, France

[2] UMR GAEL - lemarie@grenoble.inra.fr
INRA-Université Pierre Mendès - Grenoble - France

[3] UMR Agronomie - Muriel.Morison@grignon.inra.fr
INRA-Agro Paris Tech - Thiverval-Grignon -France

Résumé : L'agriculture intégrée consiste à adapter les pratiques agricoles pour limiter l'usage des produits phytosanitaires tout en contenant la pression des biogresseurs et avec une productivité suffisante pour garantir un revenu satisfaisant à l'agriculteur. A l'échelle de l'année culturale, le problème peut être formulé comme un problème d'optimisation de décisions d'un itinéraire technique (ITK). Nous avons mis en oeuvre une méthode d'apprentissage par renforcement pour adopter une méthode d'optimisation basée sur l'expérimentation virtuelle dans le cadre de l'optimisation d'ITK pour la conduite de culture du colza. Ce projet a requis le couplage de plusieurs modèles mécanistes de simulation. Des résultats préliminaires de l'optimisation sont présentés. **Mots-clés** : Apprentissage par renforcement, agronomie, conduite de culture, colza

1 Introduction

La production agricole doit répondre à des enjeux importants de réduction de l'usage d'intrants (en particulier pesticides) tout en maintenant un bon niveau de productivité. L'agriculture intégrée est une voie intéressante pour atteindre cet objectif. Son principe consiste à adapter les pratiques agricoles pour limiter l'usage des produits phytosanitaires tout en contenant la pression des biogresseurs et avec une productivité suffisante pour garantir un revenu satisfaisant à l'agriculteur. Cette adaptation peut se faire à deux niveaux i) sur la rotation culturale (succession de cultures au cours du temps sur la parcelle) ii) sur l'itinéraire technique (la succession des opérations culturales : labour, semis ...) d'une culture donnée. Les recherches agronomiques dans ce domaine s'appuient d'une part sur les expérimentations aux champs mais aussi sur l'expérimentation virtuelle par simulation informatique.

La modélisation des cultures visant à concevoir et comparer de nouveaux itinéraires techniques intégrés a fait l'objet d'une littérature importante en agronomie. D'une manière générale, ces travaux consistent à comparer différents itinéraires techniques définis a priori par des agronomes en les simulant avec un modèle agronomique pour des conditions pédoclimatiques données. L'itinéraire technique recherché est celui qui maximise une fonction objectif (ex : espérance de marge brute). En économie agricole, de nombreux travaux se sont appuyés sur une modélisation bioéconomique pour explorer plus en détail les meilleures solutions qui pourraient être proposées Sexton *et al.* (2007). Le présent article entre dans cette tradition d'optimisation d'un modèle bioéconomique. L'optimisation porte sur plusieurs décisions d'un itinéraire technique, prises à différents moments dans le cycle de production. L'analyse est appliquée au cas du colza pour lequel nous disposons d'un modèle de culture assez complet (OmegaSys) intégrant l'effet des bioagresseurs. Compte tenu de la complexité du modèle qui est optimisé, la résolution de ce programme dynamique repose sur la méthode proposée dans Garcia & Ndiaye (1998) qui est un algorithme d'apprentissage par renforcement Sutton & Barto (1998).

La programmation dynamique permet d'optimiser une séquence de décisions pour le pilotage d'un système dynamique aléatoire Kennedy (1986). Les aléas sont présents tout au long du cycle de production

et sont dus dans notre cas au climat. En conséquence, certaines décisions de l'itinéraire technique sont postérieures à certains aléas. Il est alors généralement préférable de ne pas fixer toutes les décisions de l'itinéraire technique, mais de laisser la possibilité d'adapter certaines d'entre elles à l'état de la culture au moment de prendre la décision. Ce principe d'adaptation des décisions est bien pris en compte par les agronomes lorsqu'ils définissent des règles de décision dans leurs itinéraires techniques. Néanmoins ces règles de décisions sont définies a priori. Dans cet article, nous relâchons ces contraintes en ne définissant que les variables sur lesquelles doivent s'appliquer les décisions, sans définir de règles a priori.

Le modèle agronomique de la culture du colza qui est utilisé ici est relativement complexe car il doit permettre de prévoir les effets d'un nombre important de décisions de l'itinéraire technique. Un tel modèle possède de nombreuses variables d'état continues. La solution qui est retenue ici consiste à s'appuyer sur un algorithme d'apprentissage par renforcement qui permet d'approcher par le biais de la simulation les solutions exactes du problème. L'implémentation informatique du projet a été réalisée sur la plateforme RECORD Bergez *et al.* (2013) qui permet de combiner différents modules indépendants, ce qui permet de séparer la programmation des modèle agronomiques, climatique et décisionnelle. Cette plateforme permet d'intégrer facilement ces différents modules et de les recombiner avec d'autres modules à terme si nécessaire.

La section suivante est consacrée à une présentation plus détaillée des problèmes posés par le raisonnement des itinéraires techniques intégrés, et la façon dont ces problèmes ont été traités dans la littérature en agronomie et en économie agricole. Nous présenterons ensuite l'ensemble du modèle bioéconomique comprenant le module d'optimisation (section 3). Des résultats préliminaires sont présentés dans la section 4, avant de conclure et de proposer des futurs travaux.

2 Problématique et litterature

Un itinéraire technique est défini comme l'ensemble des opérations techniques réalisées au cours d'un cycle de production : travail du sol, semis (densité et variété), apport azoté, traitements phytosanitaires. Pour chacune des ces opérations l'agriculteur choisit à la fois une option technique et une date de réalisation, certains de ces choix étant néanmoins contraints par la disponibilité de l'agriculteur ou la possibilité de faire une intervention. Pour concevoir de nouveaux itinéraires techniques, les agronomes doivent prendre en compte la dimension systémique de la culture qui est conduite. Les différentes décisions de l'itinéraire technique interagissent entre elles si bien qu'il n'est généralement pas possible d'optimiser ces décisions de façon indépendante. Le nombre de combinatoires de décisions étant généralement très large, la conception et la validation de nouveaux itinéraires techniques soulève des problèmes méthodologiques importants. Une autre difficulté tient au fait que l'impact d'un itinéraire technique peut être différent selon le climat de l'année, climat que l'agriculteur ne connaît pas parfaitement à l'avance. Il convient donc de pouvoir dégager des itinéraires techniques qui soient intéressants sur une gamme assez large de climats représentatifs d'une certaine région.

D'une manière générale, trois grands types de méthodes sont utilisés et combinés par les agronomes pour concevoir de nouveaux itinéraires techniques répondant à plusieurs objectifs agronomiques, économiques et environnementaux. La première s'appuie sur le prototypage Lançon *et al.* (2007); Rapidel *et al.* (2009) : il s'agit alors de mobiliser plusieurs acteurs et de concevoir des systèmes prototypes à partir de leur connaissance experte, ces prototypes étant ensuite évalués par des expérimentations au champ coûteuses. La deuxième méthode procède "pas à pas" par des aller-retour entre la conception et l'évaluation au champ en mobilisant peu d'experts mais en s'appuyant sur les connaissances agronomiques acquises par ailleurs et par un schéma de fonctionnement de la culture. Enfin, la troisième méthode s'appuie sur la réalisation de simulations soit à partir de modèles simples (bilan d'azote, bilan hydrique, etc) ou complexes (modèle de cultures ; David *et al.* (2004); Van Ittersum *et al.* (2003)) ou des modèles spécifiquement construits pour classer et identifier les systèmes de culture les mieux adaptés Loyce *et al.* (2002); Aubry *et al.* (1998).

Le recours à l'usage de modèles agronomiques présente plusieurs avantages. Tout d'abord, ils permettent de simuler le fonctionnement de la culture avec différents sols et différents climats. Les résultats permettent de dégager les jeux de décisions qui sont les plus adaptés aux différents contextes pédoclimatiques possibles pour la culture. Certaines de ces décisions peuvent être des règles qui prennent en compte l'état de la culture au moment où la décision doit être réalisée. En second lieu, un modèle agronomique assez complet permet de simuler les effets de différentes décisions de l'agriculteur. La dimension systémique est bien prise en compte et les interactions entre les pratiques agricoles peuvent être étudiées. Il est alors possible

de dégager le meilleur ensemble de décisions pour répondre à un certain objectif dans un contexte donné. Enfin, lorsque plusieurs objectifs sont poursuivis, l'emploi du modèle peut aider à établir des compromis entre ces objectifs. Par exemple, en quoi un poids plus important mis sur l'objectif 1 facilite ou compromet la réalisation de l'objectif 2 ?

Une première limite de ces travaux de modélisation porte sur la façon dont les décisions sont définies et optimisées. Rappelons que la difficulté méthodologique est de taille puisqu'il s'agit d'optimiser un ensemble de décision prises de manière séquentielle dans un contexte aléatoire avec des interactions entre ces décisions. Il est utile ici de confronter les travaux réalisés en agronomie d'un coté et en économie de l'autre. Du coté des agronomes, il existe maintenant un certain nombre de travaux qui raisonnent des ensembles de décisions en prenant en compte leurs interactions. La nécessité de s'adapter aux effets des aléas tout au long du cycle de production est prise en compte lorsque les agronomes définissent des règles de décisions. Néanmoins, ces règles de décisions Chatelin *et al.* (2005) ou ces options techniques Loyce *et al.* (2002) sont fixes et définies par l'expertise de l'agronome. En d'autres termes, l'agronome utilise le modèle pour simuler et comparer l'effet des jeux de décisions qu'il juge intéressant, mais le modèle sur lequel il s'appuie ne permet pas directement de générer ou d'explorer de nouveaux jeux de décision.

En économie agricole, les premiers travaux sur l'optimisation d'une pratique agricole ont concerné des cas assez simples sur une décision donnée. A la différence de ce qui a été écrit plus haut à propos des modèles agronomiques, la recherche de la solution optimale fait partie intégrante du modèle. De nombreux travaux ont été réalisés sur la question de la lutte intégrée contre les bioagresseurs (*Integrated Pest Management*). L'objectif est alors de définir une règle sur la décision ou non de traiter contre un bioagresseur donné. La notion de seuil d'intervention a été proposée initialement par Headley (1972) et de nombreuses applications ont été réalisées ensuite sur différentes cultures en prenant en compte l'incertitude sur les attaques de pathogènes et l'aversion au risque de l'agriculteur. Le calcul d'un tel seuil d'intervention ne pose pas de problème méthodologique majeur dès lors qu'on dispose d'une fonction de dommage établissant le lien entre la densité de bioagresseur et la perte de rendement qui en résulte. L'extension de ce travail à des combinaisons de décisions ne pose pas de problème majeur lorsque ces décisions sont prises simultanément.

L'optimisation d'un ensemble de décisions séquentiel en présence d'aléas a fait l'objet de beaucoup moins de travaux. Le problème posé ici est un problème de programmation dynamique Kennedy (1986). La modélisation de l'effet des différentes décisions nécessite de développer des modèles agronomiques assez complexes. Garcia (1999) et Bergez *et al.* (2001) proposent d'utiliser des méthodes de résolution de processus décisionnels de Markov (PDM) pour l'optimisation d'ITK. Il s'agit respectivement d'optimiser les apports azotés sur blé au cours du cycle de production et les pratiques d'irrigation pour le maïs. La résolution directe de PDM par programmation dynamique n'étant pas possible, du fait d'une part de la présence de variables d'états continues et d'autre part de la non disponibilité des probabilités de transitions du PDM, les auteurs ont recours à des algorithmes d'apprentissage par renforcement. Le présent article est très proche de ces deux travaux puisque l'optimisation est faite en utilisant le même type d'algorithme. L'originalité principale tient au fait que la culture étudiée est le colza, que certaines des décisions considérés portent sur des traitements pesticides et également que l'implémentation est faite ici sur une plateforme de simulation (RECORD) qui permet de combiner différents modules élémentaires. Depuis, à notre connaissance, aucune application de l'optimisation d'ITK dans un cadre PDM n'a été proposée ; cependant, il existe encore un intérêt dans ce type d'approches ; par exemple, Dimokas *et al.* (2009). On note également des travaux plus récents dans le cadre PDM, pour le domaine appliqué de la préservation d'espèces en écologie, Nicol & Chadès (2011).

3 Le modèle bioéconomique

Le modèle bioéconomique présenté ici comprend quatre parties principales représentées dans la figure 1 : le modèle agronomique, le modèle économique, le modèle décisionnel et le modèle climatique. Les modèles agronomiques et décisionnels sont les parties les plus compliquées et seront décrites en détail ci-dessous. Le modèle décisionnel définit les décisions de l'itinéraire technique et réalise l'apprentissage qui permet d'optimiser les décisions. Le modèle agronomique prend les décisions et le climat en entrée et génère un rendement final en sortie. Le modèle économique calcule la marge brute à partir des décisions prises et du rendement. Le modèle climatique fournit des séries climatiques (réelles ou simulées) en entrée du modèle agronomique et correspond à la seule source d'aléas dans le modèle considéré ici.

Il est à noter que, dans la version actuelle du modèle, le modèle phoma, qui sera décrit plus loin, présente

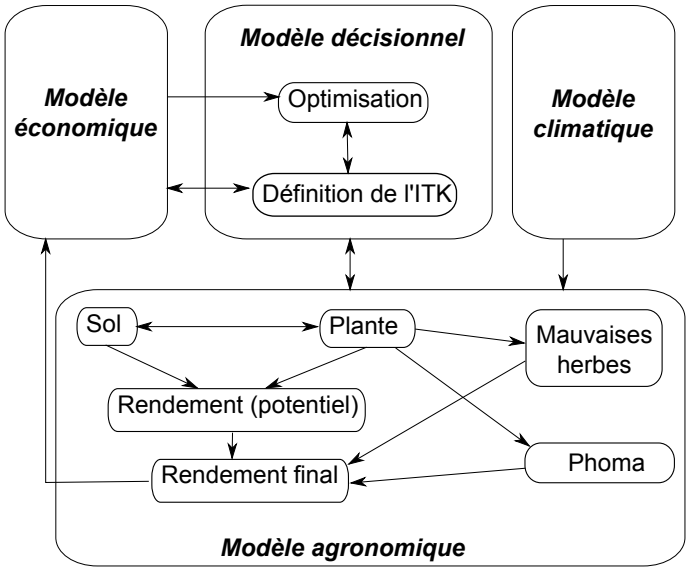

FIGURE 1 – Représentation synthétique du modèle bioéconomique

certains problèmes qui font que la décision de traitement contre le phoma n'a pas d'impact sur le rendement. Le modèle phoma sera décrit plus loin. Nous l'ignorerons ensuite dans la présentation du modèle décision et des résultats.

3.1 Le modèle agronomique

OMEGAS sys (*Oil seed rape Model to Experiment and Generate Alternative Systems*) est un modèle biotechnique qui simule de manière dynamique la croissance de la culture de colza sous l'effet de facteurs limitant abiotiques (eau et azote) et les pertes de rendement liées à deux bioagresseurs, les mauvaises herbes et le phoma. Omegasys correspond au module agronomique de la figure 1. Il se compose d'un sous-module de la dynamique de la plante, et de deux sous-modules de bioagresseur (Phoma et Mauvaises herbes). Le sous-module plante simule jour après jour la croissance en biomasse, le LAI (*Leaf Area Index* qui représente la part de la surface folliaire par rapport à la surface au sol), l'absorption d'azote et les carences azotées de la culture, la croissance racinaire et la dynamique d'eau dans le sol. Il calcule en fin de cycle le rendement potentiel de la culture sous l'influence cumulée sur l'année des carences azotées et du déficit hydrique. Ce modèle est très largement dérivé d'un modèle de culture dynamique, Azodyn Colza Jeuffroy *et al.* (2006); Valantin-Morison *et al.* (2004). Ce sous-module est combiné à un sous-module statique estimant la biomasse de mauvaises herbes à l'entrée de hiver, en se basant sur les travaux de Primot *et al.* (2006), et d'un sous-module phoma, estimant l'intensité de l'infection primaire (nombre de macules), la gravité de la maladie en fin de cycle Aubertot *et al.* (2004). Ces deux sous-modules impactent en fin de cycle le rendement potentiel fourni par le module de plante. Les variables d'entrées sont de quatre catégories : (1) Variables initiales concernant les mauvaises herbes : espèces et densité de mauvaises herbes à la levée ; (2) celles concernant le phoma : inoculum primaire (nombre de macules) (3) les variables de l'itinéraire technique, elles-mêmes objet de décision et en lien avec le module décisionnel (voir section 3.3) : date de semis, densité de semis, travail du sol, résistance variétale, fongicide herbicide, fertilisation azotée, écartement ; les variables pédo-climatiques : teneur en argile, calcaire, teneur en azote organique, densité apparente, épaisseur de minéralisation, amendements organiques et résidus des années précédentes, contenu en eau et en azote au début de la simulation et les données climatiques (rayonnement, évapo-transpiration, pluie, températures minimales et maximales).

Les sorties du modèle sont de trois types : (1) le rendement potentiel, la teneur en huile, le rendement après l'effet compétition des mauvaises herbes, et l'attaque du phoma (2) la gravité de la maladie et la biomasse de mauvaises herbes (3) la quantité d'azote lixivié sous la culture pendant tout le cycle et la quantité d'azote contenue dans le sol à la récolte. Dans le cadre de ce travail nous ne mobiliserons que les variables de rendement potentiels et les rendements affectés par les bioagresseurs.

3.2 Le modèle climatique

Les données climatiques constituent ici les paramètres aléatoires du modèle, ce qui fait du problème d'optimisation un problème de décision séquentielle dans l'incertain. Les entrées climatiques du modèle agronomique sont des données journalières de rayonnement global, de températures, de précipitations et d'évapo-transpiration (ETP). Deux alternatives sont possibles. Dans ce papier, seule la première a été mise en oeuvre mais l'objectif est de se reposer sur la deuxième (voir section 4.2).

La première, de type échantillonnage parmis un ensemble de séries climatiques observées, consiste à tirer aléatoirement en début de simulation une des 41 séries climatiques disponibles pour la simulation qui sont issues des stations météorologique de Versailles et Grignon, entre 1971 et 2008.

La seconde alternative, de type génération de série climatique pseudo-aléatoires, repose sur l'utilisation d'un générateur stochastique de données météorologiques.

3.3 L'itinéraire technique du modèle décisionnel

Les 4 décisions de l'itinéraire technique (ITK) considérées sont : la date de semis, l'apport d'azote à l'automne et au printemps et le traitement herbicide. Ces décisions sont décrites dans le tableau 1.

TABLE 1 – Paramètres de l'itinéraire technique

Décision (date de prise de décision)	Modalité technique et date de réalisation	Variables observées
Semis (30/07)	0) 01/08 1) 01/09 2) 15/09	- Azote disponible dans le sol (Nsol) - Eau disponible dans le (stockC2apdrainage)
Azote automne (25 jours après la levée)	0) 40 unités le lendemain 1) 100 unités le lendemain	- Azote disponible dans le sol (Nsol) - Eau disponible dans le (stockC2apdrainage) - Matière sèche total générée (MSTg) - Indice de nutrition (INN) - Décision semis (DecSemis)
Traitement herbicides (30 jours après la levée)	0) aucun traitement 1) traitement le lendemain	- Azote disponible dans le sol (Nsol) - Bilan hydrique pour les mauvaises herbes (HydriqueMH) - Matière sèche total générée (MSTg) - Decision semis (DecSemis) - Décision azote automne
Azote printemps (13/02)	0) 40 unités le 01/03 1) 33 le 25/02 + 66 le 05/03 2) 50 le 23/02 + 50 le 02/03 + 50 le 09/03 3) 40 unités le 01/04 4) 33 le 25/03 + 66 le 05/04 5) 50 le 23/03 + 50 le 02/04 + 50 le 09/04	- Azote disponible dans le sol (Nsol) - Matière sèche totale générée (MSTg) - Decision semis (DecSemis) - Decision azote automne (DecAzoteAutomne) - Decision herbicides (DecHerbicides)

Il est important de différencier les dates de prise de décision des dates de réalisation de ces décisions. Les dates de prise de décision permettent de définir la séquence des décisions qui peut être représentée par exemple avec un arbre de décision. Les dates de réalisation des décisions sont celles qui ponctuent les moments où le modèle agronomique est affecté par l'itinéraire technique. Lorsque plusieurs dates de réalisation sont possibles comme c'est le cas avec le semis ou l'apport d'azote au printemps, ces différentes dates représentent différentes branches de l'arbre de décision. Les dates auxquelles les deux décisions d'automne post-semis sont prises sont exprimées en relatif à la date de levée de la culture. Ceci assure que, quelle que soit la date de semis, ces décisions post-semis soient prises et réalisées à des stades équivalents de la culture.

Les périodes de décision sont représentées sur un axe temporel dans la figure 2. Le modèle économique intègre au cours de la simulation les différents coûts liés aux actes techniques ainsi que le gain au moment de la récolte, fonction du rendement. Nous nous plaçons ici dans un cadre où les séquences de décisions

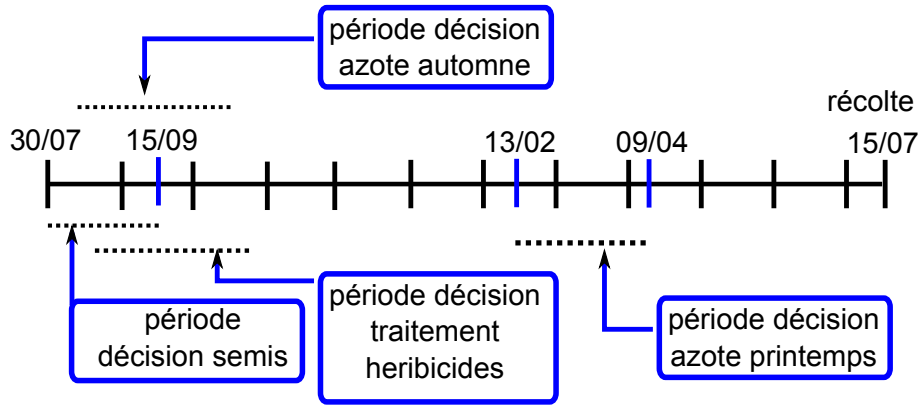

FIGURE 2 – Les différentes périodes de décision représentées (approximativement) sur un axe temporel

sont clairement distinctes : la date de réalisation de la nième décision précède toujours la date de prise de la n+1ième décision. De plus, les coûts de chacune des décisions sont intégrés avant la prise de décision suivante, ce qui est une contrainte qui doit être satisfaite pour la mise en œuvre de l'optimisation.

Comme nous le verrons dans la section suivante qui présente le programme d'optimisation, chaque décision est prise en fonction d'un certain nombre de variables observées le jour de la prise de décision. Les variables observées qui ont été retenue pour notre analyse sont présentées dans le tableau 1. Les variables sont pour la plupart des variables continues caractérisant l'état de la plante ou du sol, données par le modèle agronomique au moment de la prise de décision.

3.4 Le module d'optimisation du modèle décisionnel

3.4.1 Le cadre général : PDM et programmation dynamique

L'optimisation de décisions séquentielles peut se formuler dans le cadre des processus décisionnels de Markov (PDM) et résolue par programmation dynamique Puterman (1994). Un PDM est un graphe où les nœuds représentent les états du système et les arcs les transitions entres ces états. Une politique est une fonction $\pi : S \rightarrow D$ qui associe à un état $s \in S$ une décision à appliquer $d \in D$. L'application de cette décision résulte en une transition vers un autre état s' et la réception d'une "récompense". On note $p(s'|s,d)$ la probabilité d'atteindre l'état s' depuis l'état s en appliquant la décision d. L'hypothèse de Markov est vérifiée si $p(s'|s^n,d^n) = p(s'|s^n,d^n,s^{n-1},d^{n-1},\ldots,s^0,d^0)$ où $s^n,d^n,s^{n-1},d^{n-1},\ldots,s^0,d^0$ représente l'historique des états visités et des décisions associées. On note également $r(s,d,s')$ la récompense associée à la prise de décision d depuis l'état s avec comme état d'arrivée s'.

Le problème d'optimisation d'un PDM consiste à construire la politique π^* qui permet de maximiser les récompenses reçues à long terme. Dans notre cas, l'ensemble des étapes de décision de l'itinéraire technique au cours d'une saison culturale est décidé a priori et est fini. Cela suggère l'utilisation de méthodes d'optimisation à horizon fini. De plus, puisque le critère à optimiser est la somme des coûts de désherbage (en négatif) et de la marge brute, nous avons opté pour une méthode d'optimisation sans pondération des récompenses.

Le critère que l'on cherche à maximiser est donc le gain total espéré : $\mathbb{E}[r_1 + \ldots + r_N]$ où N est le nombre d'étapes de décision au cours de la saison culturale, et où $r_1,\ldots,r_N$ représente la séquence des récompenses reçues. Il faut noter qu'il s'agit ici d'une espérance car le modèle est stochastique par le fait en particulier de l'utilisation de données climatiques en entrées.

Pour résoudre ce problème d'optimisation, on se base classiquement sur l'évaluation des équations de Bellman. On considère une fonction de valeur $V_i^\pi : S_i \rightarrow \mathbb{R}$, où $i \in \{1,\ldots,N\}$ identifie l'étape de décision et S_i représente l'espace d'état pour l'étape i. $V_i^\pi(s)$ est une estimation, lorsque l'on applique la politique π, du gain total espéré des récompenses depuis l'état s : $V_i^\pi(s) = \mathbb{E}_\pi[r_i + \ldots + r_N | s_i = s]$ où s_i est l'état à l'étape i. La programmation dynamique consiste alors à déterminer les fonctions de valeur optimales V_i^*, solutions des équations suivantes (avec $V_{N+1}^* = 0$) :

$$V_i^*(s) = \max_{d\ in D} \sum_{s' \in S_{i+1}} p(s'|s,d) \times (r(s,d,s') + V_{i+1}^*(s'))$$

En fait, cette équation une fois résolue permet d'obtenir une relation directe entre V_i^* et les états de l'étape suivante S_{i+1}. Les politiques π_i^* sont simplement définies de la manière suivante :

$$\pi_i^*(s) = \underset{d\ inD}{\operatorname{argmax}} \sum_{s' \in S_{i+1}} p(s'|s,d) \times (r(s,d,s') + V_{i+1}^*(s'))$$

3.4.2 Apprentissage par renforcement à horizon fini et sans pondération

La programmation dynamique requiert la connaissance du PDM pour lequel on recherche une politique optimale et estimer les probabilités de transition entre états $p(s'|s,d)$ demande une connaissance fine du système étudié. Lorsque ce système est trop complexe, les méthodes d'apprentissage par renforcement permettent de s'affranchir de cela. Souvent, elles requièrent seulement la possibilité de pouvoir simuler ces transitions $s, d \rightarrow s', r$. Toutefois, l'hypothèse de processus markovien doit toujours être satisfaite. Schématiquement, par analogie avec l'apprentissage par essais-erreur, on peut représenter le principe de l'apprentissage par renforcement par une interaction entre un agent apprenant et un environnement : l'environnement fournit une observation s à l'agent qui lui-même fournit une action d à l'environnement, ce dernier simule la transition $s, d \rightarrow s', r$ et fournit à l'agent à la fois une récompense r liée à cette transition et un nouvel état s'. En itérant un nombre important de fois cette expérience, les méthodes d'apprentissage par renforcement cherchent à construire une politique optimale. A partir de ces transitions observées, on cherche en général (par exemple Watkins (1989)) à mettre à jour non plus les valeurs d'états V_i^π de la programmation dynamique mais des Q valeurs : $Q_i^\pi : S_i \times D_i \rightarrow \mathbb{R}$, $Q_i^\pi(s,d)$ estime le gain total espéré en appliquant la décision $d \in D_i$ puis en appliquant par la suite la politique π. A partir de données issues d'une transition, la mise à jour des Q valeurs peut se faire de la manière suivante :

$$Q_i(s,d) \leftarrow Q_i(s,d) + \alpha \times [r + \max_{d' \in D_{i+1}} Q_{i+1}(s',d') - Q_i(s,d)]$$

Où $\alpha \in [0;1[$ représente un taux de mise à jour qui doit converger vers 0 au cours de l'apprentissage. Cet algorithme converge vers la politique optimale lorsque l'ensemble des paires (s,d) est continuellement exploré. Par rapport à la programmation dynamique, cela requiert que l'agent applique une politique d'exploration de cet espace. En première approche nous avons utilisé la politique $\epsilon-$*greedy*, avec $\epsilon \in [0;1]$, qui consiste à chaque étape à choisir la décision préconisée par la politique courante $\pi(s) = \operatorname{argmax}_{d\ inD} Q(s,d)$ avec une probabilité $1-\epsilon$ et à choisir une autre action aléatoirement avec une probabilité ϵ.

Pour le cas particulier de l'horizon fini sans pondération des récompenses, nous avons opté pour l'algorithme R_H*-learning* Garcia & Ndiaye (1998) (voir algorithme 1) qui estime, pour la politique optimale, la valeur moyenne des récompenses obtenues à chaque étape $\rho = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} r_i]$ ainsi que les valeurs relatives à cette moyenne pour chacune des étapes :

$$R_i(s,d) = \mathbb{E}[\sum_{j=1}^{N} r_j - \rho | s_i = s, d_i = d]$$

Où $s_i \in S_i$ et $d_i \in D_i$ sont respectivement un état et une décision de départ pour l'étape $i \in \{1,\ldots,N\}$. La politique apprise est obtenue en sélectionnant les décisions qui maximisent ces valeurs relatives.

$$\forall i \in \{1,\ldots,N\}, \forall s \in S_i, \pi_i(s) = \underset{d\ inD_i}{\operatorname{argmax}} R_i(s,d) \quad (1)$$

Puisque les domaines S_i sont continus, des méthodes de régression sont utilisées de manière à généraliser les transitions et récompenses observées pendant la simulation. Les valeurs relatives $R_i(s,d)$ sont sous la forme de combinaisons linéaires de p_i fonctions caractéristiques (notées $\phi_{i,k}$ pour $k \in \{1,\ldots,\pi\}$) :

$$R_i(s,d) = \sum_{k=1}^{p_i} \omega_{i,k} \times \phi_{i,k}$$

L'apprentissage consiste alors à estimer l'ensemble des paramètres $\omega_{i,k}$. Dans l'algorithme R_H*-learning* présenté ci-dessous, les fonctions $R_i(s,d)$ sont mises à jour de manière incrémentale en fonction de l'erreur de prédiction de la récompense immédiate. La méthode repose sur des régression de type CMAC, équivalent au *Tile Coding* Sutton & Barto (1998). Celles-ci sont basées sur des fonctions caractéristiques $\phi_{i,k}$ binaires, c'est-à-dire qu'elles prennent la valeur 1 sur un sous domaine de $S \times D$ et 0 sinon. De cette manière, pour un couple $(s,d) \in S \times D$, l'erreur de prédiction est utilisée pour mettre à jour les paramètres $\omega_{i,k}$ correspondant aux fonctions $\phi_{i,k}$ activées (c-à-d. pour lesquelles $\phi_{i,k}(s,d) = 1$). Les paramètres de cet algorithme (annexe A) sont α et β qui sont respectivement les taux de mise à jour des valeurs relatives

$R_i(s,d)$ qui sont respectivement les taux de mise à jour des valeurs relatives ρ. Le taux d'exploration est représenté par ϵ. Il faut également spécifier un nombre maximal de simulations et spécifier les paramètres propres aux régressions CMAC, à savoir le nombre de grille et le pas de discrétisation, qui vont déterminer le nombre de paramètres $\omega_{i,k}$ à estimer par l'apprentissage.

Algorithme 1 : L'algorithme R_H*-learning*

initialiser $\omega_{i,k} = 0$;
initialiser $\rho = 0$;
initialiser le nombre de simulations effectuées $m = 0$;
pour $m \in 1, \ldots, N$ **faire**
  initialiser l'étape de décision : $n = 1$;
  simuler jusque la première étape de décision et observer l'état s_1;
  choisir la décision d_1 selon une politique ϵ-*greedy*;
  tant que $n < N$ **faire**
    simuler jusque la prochaine étape et observer l'état s_{n+1};
    observer une récompense associée r_n;
    erreur de prédiction immédiate : $e_n = r_n - \rho + \max_{d \in D_{n+1}} R_{n+1}(s_{n+1}, d) - R_n(s_n, d)$;
    mettre à jour la valeur relative en modifiant les $\omega_{n,k}$: $R_n(s_n, d_n) \leftarrow R_n(s_n, d_n) + \alpha \times e_n$;
    si d_n *est la décision optimale selon la poltique courante* **alors**
      mettre à jour : $\rho \leftarrow \rho + \beta \times e_n$;
    fin
    choisr la décision d_{n+1} selon une politique ϵ-*greedy*;
    $n = n + 1$;
  fin
  simuler jusque la fin et observer la récompense r_N;
  calculer l'erreur de prédiction immédiate : $e_N = r_N - \rho + R_N(s_N, d)$;
  mettre à jour la valeur relative, en modifiant les $\omega_{N,k}$: $R_N(s_N, d_N) \leftarrow R_N(s_N, d_N) + \alpha \times e_N$;
  si d_N *est la décision optimale selon la poltique courante* **alors**
    mettre à jour : $\rho \leftarrow \rho + \beta \times e_N$;
  fin
fin

3.5 Implémentation sur la plateforme RECORD

Ce projet a été développé suivant une approche systémique qui a pour avantage de fournir un cadre méthodologique pour la modélisation conceptuelle de systèmes complexes intégrant plusieurs disciplines. Ainsi, les sous-systèmes ont été spécifiés, et ensuite implémentés informatiquement en autant de modèles et sous modèles sur la plate-forme RECORD (les sous-systèmes sont organisés de manière hiérarchique). Les interactions entre ces sous systèmes ont été établies, et ensuite implémentées en autant de flux d'information transitant par les ports d'entrée et de sortie des modèles. La plate-forme garantit la cohérence des couplages entre les différents modèles et offre des services facilitant le paramétrage et le pilotage des simulations.

Comme il a été décrit dans la figure 1, le modèle a été décomposé en 4 sous-systèmes principaux : le modèle agronomique, le modèle décisionnel, le modèle climatique, et le modèle économique. Chacun de ces sous-systèmes a été traité de manière indépendante, conformément à la représentation systémique communément appelée approche "boîte noire". Le modèle agronomique a été implémenté suivant le modèle conceptuel Omega-Sys (cf. section 3.1) dans le formalisme mathématique des équations aux différences. Le modèle décisionnel "Decision" (cf. section 3.3) a été créé sur la base des activités culturales liées à la conduite du colza, dans le formalisme proposé par la plate-forme.

Les quatre sous-systèmes échangent au cours de la simulation différentes informations. Le modèle climat envoie à chaque pas de temps les variables climatiques au modèle agronomique. Le modèle décisionnel reçoit à chaque pas de temps les valeurs des variables d'état d'intérêt pour décider des différentes activités culturales, ces valeurs proviennent du modèle agronomique (Indice de nutrition azotée, azote du sol ...) et du modèle économique (récompense) en lien avec l'algorithme d'apprentissage. Le modèle agronomique reçoit de manière événementielle, les informations issues du modèle décisionnel et susceptibles de modifier sa dynamique à savoir les opérations de semis, d'apport d'azote avec la quantité d'azote associée et de

récolte.

4 Résultats préliminaires

Dans un premier temps, il nous a semblé pertinent d'étudier le besoin de recourir à la construction de stratégies adaptatives plutôt que de proposer des itinéraires techniques fixes, donnés a priori. Nous avons donc procédé aux simulations de l'ensemble des itinéraires techniques, en combinant toutes les décisions (section 3.3) avec chacune séries climatiques disponibles. Un ITK est identifié par 4 chiffres. Par exemple, l'ITK 2103 correspond à un semis le 15/09 (décision 2), un apport d'azote en automne de 100 unités (décision 1), pas d'apport d'herbicides (décision 0) et un apport d'azote de 40 unités le 01/04 (décision 3).

La figure 3(à gauche et au centre) récapitule les résultats issus de ces simulations. A gauche, les quantiles du profit sont représentés pour chacun des ITK, et montre une variabilité importante liée à l'aléa climatique. Si quelques un des ITK semblent plus efficaces pour optimiser le profit, il n'est pas forcément évident de choisir celui qui est robuste à l'aléa climatique. En l'occurrence, l'ITK a priori qui maximise le profit est l'ITK 1000, il permet d'obtenir un profit moyen de 416 euros. A droite, on retrouve le nombre de séries climatiques pour lesquelles un ITK maximise le profit. On remarque une diversité dans les ITK optimaux. On remarque également que l'ITK 0000 est celui qui, le plus souvent, maximise le profit (sur 13 séries climatiques), bien que l'ITK optimal a priori soit l'ITK 1000 (optimal pour 11 séries climatiques). Ces résultats peuvent être comparés avec ceux obtenus si l'on prenait le meilleure ITK (*a posteriori*) en fonction de chaque climat. Dans ce cas, le profit moyen optimal est de 492. On peut donc conclure, sur l'ensemble de ces séries climatiques, que la marge de progression en adaptant la stratégie est de 76 euros.

Pour l'apprentissage, nous avons fixé le paramètre β à 0.1 et le paramètre α à la valeur importante de 1, de manière à procéder à une intialisation rapide des $\omega_{i,k}$ lors de l'apprentissage. Les fonctions de décroissance sont respectivement $\alpha_n(s,d) = \alpha/(1 + log(N(s,d)))$, où $N(s,d)$ est le nombre d'occurences du couple $s * d$ et $\beta_n = \beta/(1 + log(n))$. Le paramètre d'exploration ϵ a été fixé à 0.4. Il faut également spécifier, pour chacune des étapes de décision les paramètres propres aux modèles de régression CMAC, qui vont déterminer le nombre de paramètres $\omega_{i,k}$. Cependant seuls les paramètres correspondant aux régions les plus explorées par l'apprentissage devront être estimés précisément. Nous avons respectivement, pour les étapes de décision du semis, de l'apport d'azote en automne, de l'apport d'herbicides et de l'apport d'azote au printemps utilisés 3 grilles avec une discrétisation de 20, 3 grilles avec une discrétisation de 10, 3 grilles avec une discrétisation de 10 et 2 grilles avec une discrétisation de 15. Nous avons effectué 500 000 simulations, de manière à s'assurer que les estimations aient convergé (environ 3 jours sur un Intel Xeon 2.27GHz).

4.1 Expérimentations à partir d'une base de séries climatiques

Pour comparaison avec les ITK optimaux présentés dans la figure ci dessus, la figure 3 (à droite) présente la fréquence de préconisation des ITK par la méthode d'apprentissage sur l'ensemble des séries climatiques. De manière générale on retrouve les mêmes tendances que sur les ITK optimaux (figure au centre). En particulier on retrouve souvent les ITK 0000 et 1000. On remarque de même que, pour les ITK optimaux, les apports au printemps sont souvent réduits (décisions xxx0 et xxx3). Dans les ITK préconisés, on retrouve par contre beaucoup plus souvent la décision xxx3, au dépends de la décision xxx0. Cela signifie que les apports au printemps préconisés sont plus tardifs que pour les ITK optimaux. En terme de profits, ces décisions sont très proches des ITK optimaux.

Le profit moyenné sur l'ensemble des séries climatiques des ITK préconisés par la politique construite est de 484 euros. Ce profit est proche des ITK optimaux (492 euros) et supérieur de 68 euros au profit moyen généré par un ITK optimal a priori (416 euros). En comparant les ITK qui ressortent de l'apprentissage aux ITK optimaux, on remarque que la décision d'un apport azote conséquent à l'automne (x1xx) est plus souvent sélectionné.

Les semis précoces et normaux (0xxx ou 1xxx) sont plus souvent sélectionnés que les semis tardifs et sont souvent associés à des apports de printemps réduits et l'absence de recours à l'herbicide. Cette sélection peut s'expliquer par le choix de la situation initiale : sols profonds avec reliquats d'azote autour de 40kg/ha et une faible pression de mauvaises herbes. Si l'on avait fait ces tests dans d'autres conditions de milieux et d'autres situations initiales (sols à faible teneur en eau et en azote ainsi que fortes pressions de mauvaises herbes), il est très probable que d'autres scénarios auraient été sélectionnés.

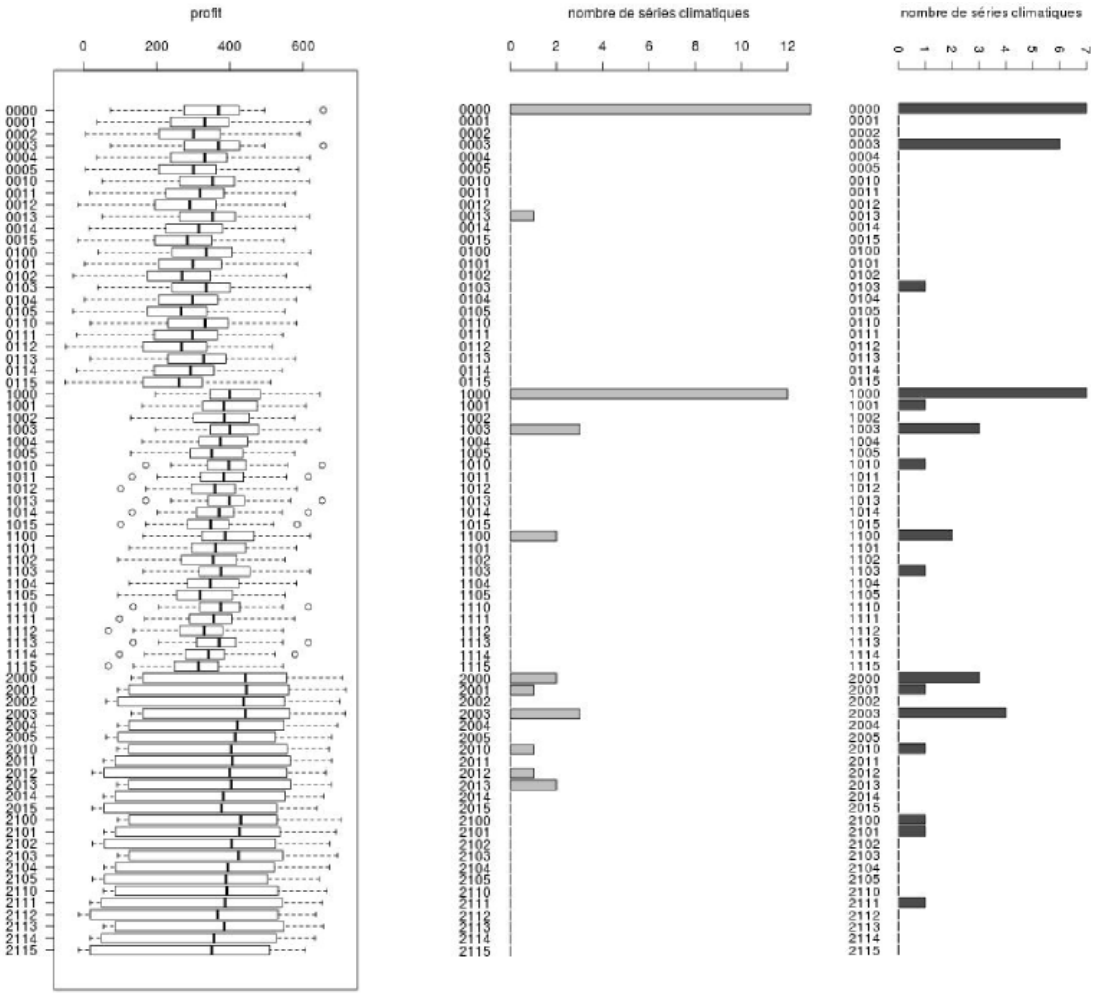

FIGURE 3 – Visualisation des résultats issus de la simulation des ITK sur la base de séries climatiques. Au gauche, les quantiles des profits simulés. Au centre, la fréquence de sélection des ITK optimaux sur la base des 41 séries. A droite, la fréquence de sélection des ITK issus de l'apprentissage sur la base des 41 séries.

4.2 Discussions sur les résultats

En premier lieu, les expérimentations préalables ont permis de montrer l'intérêt de construire des stratégies adpatatives de conduite de la culture du colza pour le semis, la fertilisation et le traitement herbicide. L'adaptation sur cet exemple a mené à une amélioration importante de la marge économique (+18%) en prenant comme référence un ITK fixé a priori à celui qui maximise l'espérance de la marge économique.

Comme nous l'avons indiqué plus haut, la décision de traitement contre le phoma n'a pas pu être prise en compte dans cette version de l'analyse car elle n'avait pas d'effet sur les résultats, mais elle reste une décision importante à prendre en compte dans ce projet.

Pour le problème reposant sur le modèle climatique qui consiste à tirer aléatoirement dans une base de séries existantes, bien que les résultats soient proches de l'optimum, le risque de sur-apprentissage est trop grand au vu de l'aléa climatique limité. Cela a permis toutefois de tester les méthodes et une politique a été construite et est éventuellement réutilisable pour le pilotage décisionnel du modèle, contrairement à une étude sur les ITKs optimaux a posteriori.

Comme mentionné dans la section 3.2, l'objectif à terme est de remplacer le modèle climatique par un générateur stochastique de données météorologiques. Un tel générateur a pour but de reproduire les propriétés statistiques des distributions des variables temporelles climatiques. Les paramètres statistiques sont induits d'une base de séries. Une fois paramétré, ce générateur peut produire une quasi infinité de séries climatiques proches statistiquement des séries observées, ce qui permet d'étudier une variabilité beaucoup plus importante de réponses du modèle aux données climatiques. Deux générateurs sont actuellement disponibles dans la plateforme RECORD. Le premier reprend la logique du générateur de série climatique WGEN Richardson (1984) qui produit les températures minimum et maximum journalières ainsi que les précipitations. Les précipitations sont modélisées suivant un processus de Markov et les températures suivant un processus d'auto-corrélation et sont conditionnées aux précipitations. La température moyenne du jour correspond à la moyenne des températures minimale et maximale. Le second, plus récent et en cours de validation, repose sur une modélisation des résidus des variables avec des distributions normales asymétriques Flecher *et al.* (2010).

Ces résultats sont donc préliminaires pour tester l'intérêt de la démarche ainsi que la faisabilité du projet, à savoir l'intégration de modèles issus de différentes communautés.

Toutefois, au vu de ces résultats, une des conclusions a été de s'orienter vers une reformulation du problème comme un problème d'optimisation continue, c'est à dire où les décisions cette fois ne sont plus discrètes mais continues. Par exemple, on ne chercherait pas à savoir laquelle des deux solutions d'apport azote automne serait optimale (40 ou 100 unités d'azote) mais on chercherait à déterminer la valeur optimale. Il faudrait alors utiliser des méthodes de regression incrémentales (*online*) qui par leur forme faciliterait l'étape d'optimisation donnée dans l'equation 1.

5 Conclusion et travaux futurs

Le problème posé dans ce papier est l'optimisation de décisions séquentielles dans le cadre de la conduite de culture du colza, avec un intérêt particulier pour l'étude des facteurs limitants (bio-agresseurs). Nous avons mis en œuvre une approche basée sur la modélisation et la simulation avec comme productions le développement et/ou le couplage d'un modèle bioéconomique reposant sur un modèle agronomique, un modèle climatique et un module d'optimisation par apprentissage par renforcement. Ces développements on été réalisés au sein de la plateforme RECORD et représentent le fruit d'une collaboration entre différentes disciplines : économie, agronomie, informatique, statistique. Nous avons présenté des résultats préliminaires qui montrent l'intérêt que peut avoir une approche basée sur la conception de stratégies de conduite adaptative. Ces expérimentations nous ont également permis de mettre en avant les travaux à conduire pour éventuellement réussir à proposer des résultats pertinents pour une interprétation agronomique. Ces travaux sont pour la plupart en cours :

— Intégration d'une étape de décision concernant la gestion du Phoma dans le modèle agronomique. Cela requiert la modélisation du traitement Phoma et de son impact économique.
— Intégration du générateur stochastique de données climatiques WACSgen Flecher *et al.* (2010). Ce générateur étant disponible dans la plateforme, des tests de paramétrisiation et de génération doivent toutefois être effectués au préalable.
— Reformuler les décisions de manière continue, et identifier des méthodes de régression plus adaptées et plus performantes.

Ces deux derniers points sont liés en terme de méthodologie et peuvent remettre en cause le type de régresseur utilisé. Nous anticipons en effet une difficulté plus importante en intégrant un générateur stochastique pour la prédiction, du fait d'un aléa plus important. En particulier, il s'agit de la prise de décision du traitement herbicide et celle de l'apport azote au printemps. Dans les deux cas, la prédiction est à long terme. Dans le 1er cas, il faut prédire l'état du 13/02 en octobre de l'année précédente, dans le 2eme cas, il faut prédire l'état du 15/07 le 13/02 (voir figure 2). Une piste à envisager serait d'utiliser d'autres méthodes de regression (Buşoniu *et al.* (2010); Vijayakumar *et al.* (2005)).

Références

AUBERTOT J.-N., PINOCHET X., REAU R. & DORÉ T. (2004). The effects of sowing date and nitrogen availability during vegetative stages on leptosphaeria maculans development on winter oilseed rape. *Crop Protection*, **23**(7), 635–645.

AUBRY C., PAPY F. & CAPILLON A. (1998). Modelling decision-making processes for annual crop management. *Agricultural Systems*, **56**(1), 45–65.

BERGEZ J., EIGENRAAM M. & GARCIA F. (2001). Comparison between dynamic programming and reinforcement learning : a case study on maize irrigation management. In *Proceedings of European Federation for Information Technology in Agriculture, Food and the Environment*.

BERGEZ J.-E., CHABRIER P., GARY C., JEUFFROY M., MAKOWSKI D., QUESNEL G., RAMAT E., RAYNAL H., ROUSSE N., WALLACH D., DEBAEKE P., DURAND P., DURU M., DURY J., FAVERDIN P., GASCUEL-ODOUX C. & GARCIA F. (2013). An open platform to build, evaluate and simulate integrated models of farming and agro-ecosystems. *Environmental Modelling and Software*, **39**(1), 39–49.

BUŞONIU L., BABUŠKA R., DE SCHUTTER B. & ERNST D. (2010). *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, Florida : CRC Press.

CHATELIN M. H., AUBRY C., POUSSIN J. C., MEYNARD J. M., MASSÉ J., VERJUX N., GATE P. & LE BRIS X. (2005). Déciblé, a software package for wheat crop management simulation. *Agricultural Systems*, **83**(1), 77–99.

DAVID C., JEUFFROY M.-H., RECOUS S. & DORSAINVIL F. (2004). Adaptation and assessment of azodyn model for managing the nitrogen fertilisation of organic winter wheat. *European Journal of Agronomy*, **21**(2), 249–266.

DIMOKAS G., TCHAMITCHIAN M. & KITTAS C. (2009). Calibration and validation of a biological model to simulate the development and production of tomatoes in mediterranean greenhouses during winter period. *Biosystems Engineering*, **103**(2), 217 – 227.

FLECHER C., NAVEAU P., ALLARD D. & BRISSON N. (2010). A stochastic daily weather generator for skewed data. *Water Ressource Research*, **46**.

GARCIA F. (1999). Use of reinforcement learning and simulation to optimise wheat crop technical management. In *Proceedings of the International Congress on Modelling and Simulation (MODSIM 99)*, p. 801–806, Hamilton, New-Zealand.

GARCIA F. & NDIAYE S. M. (1998). A learning rate analysis of reinforcement learning algorithms in finite-horizon. In *Proceedings of the 15th International Conference on Machine Learning (ML-98)*, p. 215–223 : Morgan Kaufmann.

HEADLEY J. (1972). Defining the economic threshold. *National Academy of Sciences, Pest Control Strategies for the Future*, p. 100–108.

JEUFFROY M., VALANTIN-MORISON M., CHAMPOLIVIER L. & REAU R. (2006). Azote, rendement et qualité des graines : mise au point et utilisation du modèle azodyn-colza pour améliorer les performances du colza vis-à-vis de l'azote. *Oléagineux, Corps Gras, Lipides*, **13**(6), 388–392.

KENNEDY J. (1986). *Dynamic programming : applications to agriculture and natural resources*. Elsevier Applied Science Publishers.

LANÇON J., WERY J., RAPIDEL B., ANGOKAYE M., GÉRARDEAUX E., GABOREL C., BALLO D. & FADEGNON B. (2007). An improved methodology for integrated crop management systems. *Agronomy for Sustainable Development*, **27**(2), 101–110.

LOYCE C., RELLIER J. & MEYNARD J. (2002). Management planning for winter wheat with multiple objectives (1) : The betha system. *Agricultural Systems*, **72**(1), 9–31.

NICOL S. & CHADÈS I. (2011). Beyond stochastic dynamic programming : a heuristic sampling method for optimizing conservation decisions in very large state spaces. *Methods in Ecology and Evolution*, p. 221 – 228.

PRIMOT S., VALANTIN-MORISON M. & MAKOWSKI D. (2006). Predicting the risk of weed infestation in winter oilseed rape crops. *Weed Research*, **46**(1), 22–33.

PUTERMAN M. (1994). *Markov Decision Processes*. New York : John Wiley and Sons.

RAPIDEL B., BOUBA S. T., SISSOKO F., LANÇON J. & WERY J. (2009). Experiment-based prototyping to design and assess cotton management systems in west africa. *Agronomy for Sustainable Development*, **22**(4), 545–556.

RICHARDSON C. (1984). *WGEN : A model for generating daily weather variables*. Washington, DC : National Technical Information Service (NTIS).

SEXTON S. E., LEI Z. & ZILBERMAN D. (2007). The economics of pesticides and pest control. *International Review of Environmental and Resource Economics*, **1**(3), 271–326.

SUTTON R. S. & BARTO A. G. (1998). *Introduction to Reinforcement Learning*. Cambridge, MA, USA : MIT Press.

VALANTIN-MORISON M., JEUFFROY M. & CHAMPOLIVIER L. (2004). Evaluation and sensitivity analysis of azodyn-rape, a simple model for decision support in rapeseed nitrogen. In *Congress of European Society of Agronomy (ESA'11)*.

VAN ITTERSUM M., LEFFELAAR P., VAN KEULEN H., KROPFF M., BASTIAANS L. & J. G. (2003). On approaches and applications of the wageningen crop models. *European Journal of Agronomy*, **18**(3), 201–234.

VIJAYAKUMAR S., D'SOUZA A. & SCHAAL S. (2005). Incremental online learning in high dimensions. *Neural Comput.*, **17**(12), 2602–2634.

WATKINS C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK.