Malo, Pekka; Tahvonen, Olli; Suominen, Antti; Back, Philipp; Viitasaari, Lauri

Reinforcement Learning in Optimizing Forest Management

Published in:
Canadian Journal of Forest Research

DOI:
10.1139/cjfr-2020-0447

Published: 01/10/2021

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
Reinforcement learning in optimizing forest management

Pekka Malo, Olli Tahvonen, Antti Suominen, Philipp Back, and Lauri Viitasaari

Abstract: We solve a stochastic high-dimensional optimal harvesting problem by using reinforcement learning algorithms developed for agents who learn an optimal policy in a sequential decision process through repeated experience. This approach produces optimal solutions without discretization of state and control variables. Our stand-level model includes mixed species, tree size structure, optimal harvest timing, choice between rotation and continuous cover forestry, stochasticity in stand growth, and stochasticity in the occurrence of natural disasters. The optimal solution or policy maps the system state to the set of actions, i.e., clear-cutting, thinning, or no harvest decisions as well as the intensity of thinning over tree species and size classes. The algorithm repeats the solutions for deterministic problems computed earlier with time-consuming methods. Optimal policy describes harvesting choices from any initial state and reveals how the initial thinning versus clear-cutting choice depends on the economic and ecological factors. Stochasticity in stand growth increases the diversity of species composition. Despite the high variability in natural regeneration, the optimal policy closely satisfies the certainty equivalence principle. The effect of natural disasters is similar to an increase in the interest rate, but in contrast to earlier results, this tends to change the management regime from rotation forestry to continuous cover management.

Key words: reinforcement learning, forestry, economics, stochasticity, continuous cover forestry, uneven-aged forestry, optimal rotation, mixed species, optimal harvesting.

1. Introduction

Size-structured dynamic models based on empirically realistic details of forest growth aim to develop forest management but quickly become excessively complicated when integrated with economics and optimization. Further problems arise from the need to optimize not only the rotation period but also the timing of thinning and the number of trees cut from various size classes and tree species. To obtain long-term sustainable solutions, the time horizon should be long, preferably infinite. Stochasticity is inherently involved in prices, interest rate, forest growth, and as natural disasters such as fires and windthrows. Literature offers various advanced approaches to cope with these complications, but mixed-species stochastic forestry models with detailed, realistic structures still call for further improvements. Our aim is to show that these complications can be approached by new methods developed in machine learning.

Reinforcement learning (RL) is an active branch of machine learning, with roots in psychology, neuroscience, and dynamic programming, and describes virtual agents acting in stochastic environments to maximize cumulative future rewards (Sutton and Barto 2018; Mnih et al. 2015). Modern RL-flavored algorithms have brought about astonishing scientific breakthroughs in many domains, such as robotics (Kober et al. 2013) games (Silver...
et al. 2016), economics (Charpentier et al. 2020), and bounded rationality (Leimar and McNamara 2019). RL-based methods are especially strong in solving stochastic multidimensional optimization problems that have been intractable because of the “curse of dimensionality”. Thus, they naturally correspond to problems of optimal choices in the form of sequential decision-making — a perfect match for the multidimensional forest management domain, in which harvesting decisions are made year after year based on the current forest state with the aim of maximizing cumulative revenues and (or) other objectives. To the best of our knowledge, no research exists on the potential benefits of tackling the centuries-old problem of optimal forest management (Faustmann 1849) under uncertainty with modern RL-flavored methods, although this route offers a series of benefits over existing alternatives.

We include stochasticity in stand growth and as natural disasters. Optimization models within this setup have a long history (e.g., Hool 1966). The majority of studies can be divided into two categories: those that simplify the problem by applying anticipatory solutions, i.e., do not use emerging information after the initial moment, and studies that limit the number of possible decisions and states but maintain nonanticipatory structure, i.e., produce an optimal solution (policy) as a function of stand state. Within the first category, Valsta (1992) applies a detailed single-tree model and finds the effect of growth stochasticity to be minor compared with natural disasters that shorten optimal rotation. Similar results for natural disasters are obtained in González et al. (2005). Hyttäinen and Haight (2010) solve a detailed single-tree, mixed-species model and find the fire risk to favor clearcuts compared with continuous cover management (CCF).

Within the second category, Reed (1984) and Willassen (1998) apply the generic univariate rotation model and analytically show that natural disasters shorten the rotation, while stochastic stand growth has the opposite effect. Reed and Apaloo (1991) extended the disaster result to a two-state age-biomass model and Amacher et al. (2005) include fire prevention methods. Rollin et al. (2005) specify a mixed-species size-structured Markov model with stochastic price and stand growth, describing the price by three states and the stand by 64 species or density discrete states, and solve the model by linear programming (d’Epenoux 1963). Variants of similar model structure and solution methods are later applied in Zhou et al. (2008), Buongiorno and Zhou (2015, 2017), and Zhou and Buongiorno (2019). This approach produces decision tables that map each of the 64 discrete stand states and the corresponding optimal transition to a new state depending on stochastic market prices.

Comeau and Gunn (2017) take the view that one route forward is increasing realism by including a more detailed set of state and choice variables without giving up the computation of optimal policy in the spirit of dynamic programming. For this end, they apply Neuro-Dynamic Programming. Briefly described, they apply neural network architecture in backward recursive value iterations. The procedure approximates the value function in discrete points of the state space and applies interpolants between the approximations. Comeau and Gunn (2017) compute an even-aged, two-species, eight-state-variable whole-stand model with stochastic price and natural disasters but conclude that the set of state and control variables may not yet be fully sufficient to characterize the management of real forests.

A third group of studies analyzes the financial performance of various mixed-species and stochastic setups applying portfolio theory, combinatorial models, and extensive simulation (Knöke et al. 2020; Paul et al. 2019; Messerer et al. 2020) and among other results obtain support for diversifying both forest structures and tree species.

We propose an RL-based algorithm for solving a size-structured, stochastic forestry problem that includes a detailed nonlinear mixed-species growth specification, empirical harvesting cost models, and a high number of continuous state and control variables.

We extend Parkatti and Tahvonen (2020) by including stochastic stand growth, the risk of sudden natural disasters, and by allowing any initial stand state. The problem is challenging for two reasons: first, the number of actions and states is infinite, which effectively rules out opportunities to utilize any standard method from Markov process theory versions of dynamic programming that commonly assumes either a finite number of discrete states and (or) actions. Second, to incorporate harvest timing and the choice between CCF and rotation regimes, we need an action space that is a mixture of continuous (harvested amounts) and discrete (harvest timing and type) actions.

To solve this infinite-horizon stochastic problem, we apply a new policy gradient method for RL, where we learn the optimal policy by alternating between sampling data through interaction with the environment (i.e., stand growth model) and optimizing the objective function using stochastic gradient ascent (Schulman et al. 2017). This yields an approximation of the optimal policy and the corresponding state-value function. For any stand state, we use these functions to produce a recommendation of what action to take and an estimate on the expected net present value of the given state. To ensure that the representations of policy and value functions have sufficient learning capacity while remaining differentiable, both function approximators are defined using Deep Neural Networks. The algorithm is based on an actor–critic framework, which effectively combines policy (actor) and a value-based (critic) method to solve the RL problem (Konda and Tsitsiklis 2003). In this framework, the actor constantly tries new actions (harvesting decisions) to explore the environment, while the critic’s role is to compute the advantage or relative expected benefit of taking a certain action in the given state. Learning then occurs, as the cumulated experience (i.e., observed state, action, reward trajectories) is used to update the parameters of the function approximators (neural networks) via gradient ascent (Sutton et al. 1999).

We first compute a deterministic four-tree-species size-structured problem with 44 continuous state variables, in which each variable can have an infinite number of states, and repeat the solution in Parkatti and Tahvonen (2020), who apply genetic algorithms and nonlinear programming. However, the striking difference is that the computation time with our new method is only approx. 0.06% of the computation time needed earlier.

Our method makes it natural to compute harvest solutions for initial states off the optimal trajectory starting from bare land. An already established result suggests that a higher interest rate favors CCF, but we show that given a mature stand and high fraction of large-diameter trees, a higher interest rate favors an initial clearcut although CCF would be the optimal long-term solution.

Solving a stochastic, nonlinear model with 44 continuous state and control variables together with a set of binary timing variables and with no simplifications has previously been beyond reach. The RL-based stochastic computation yields an optimal policy with the expected bare land value somewhat higher compared with the deterministic environment and reveals a standard deviation equal to 2% of the expected bare land value. Additionally, stochasticity in stand growth increases the average stand volume and the diversity of tree species composition. The high variation in natural regeneration implies that silver birch is temporarily the dominant tree species in several stochastic realizations compared with Norway spruce dominance in natural stands and in deterministic solutions with optimal harvest. Despite the high stochastic variability, the deterministic optimal policy is found to be a good approximation of the stochastic optimal policy, an outcome related to certainty equivalence. A similar feature with potentially high practical relevance has not been reported in earlier studies.

When stochasticity occurs in the form of natural disasters, the effect on optimal solution is similar to an increase in interest rate equal to the probabilistic occurrence of the disaster. This is in line with the classic result by Reed (1984), who found that the
threat of natural disasters shortens the rotation. However, our setup includes the optimization between CCF and rotation forestry (RF), and a higher interest rate favors the former. Thus, in sharp contrast with earlier results, we find that natural disasters cause a switch from RF to CCF if the former is optimal in a deterministic environment.

2. Size-structured optimization in continuous cover and rotation forestry

2.1. Deterministic specification

Before introducing the stochastic formulation of the problem, we recall the deterministic formulation by Parkatti and Tahvonen (2020) to introduce central notations that will be used throughout the paper. Thus, let the number of trees (per ha) of species $j = 1, \ldots, l$ in size classes $s = 1, \ldots, n$, at the beginning of period $t$, be denoted by $x_{jst}$. Then $x_{jst} = (x_{1st}, \ldots, x_{lstice} \in \mathbb{R}^l$ consists of the number of trees of species $j$ in all size classes and, at any time index $t$, the stand state is given by matrix $x_t \in \mathbb{R}^{l \times n}$ showing the number of trees in different species and size classes, respectively. During each period $t$, the fraction of trees of species $j$ that move from size class $s$ to the next size class $s+1$ is denoted by $0 \leq \alpha_{js}(x_t) \leq 1$. Similarly, the fraction of trees that die during period $t$ is given by $0 \leq \mu_{js}(x_t) \leq 1$. The fraction of trees of species $j$ that remain in the same size class during period $t$ equals $1 - \alpha_{js}(x_t) - \mu_{js}(x_t) \geq 0$. Natural regeneration of species $j$ is represented by the ingrowth function $\phi_j \geq 0$, with stand state $x_t$ as its argument.

Let $\Delta$ denote the period length (5 years), $r$ the interest rate per annum, and $b^s = 1/(1 + r)^s$ the discount factor. The number of harvested trees of species $j$ and size class $s$ at the end of each time period is given by $h_{jst} = \sum_{s=1}^{n} x_{jst - 1} - x_{jst}$. Where $T$ is the rotation length and $l_0$ is the time needed for artificial regeneration of trees after a clearcut. This formulation assumes that the initial state is bare land (cf. Tahvonen and Rämö 2016). Denoting by $h_t \in \mathbb{R}^{l \times n}$, the number of harvested trees in different species and size classes, the gross revenues from clear-cutting and thinning are denoted by $R_{oh}(ht)$ and $R_{th}(ht)$, and the corresponding variable clear-cutting and thinning costs are $C_{cc}(ht)$ and $C_{ct}(ht)$, respectively. We also include a fixed harvesting cost $C_T$ (planning and transporting the harvester to the site), implying that harvesting may not be carried out at each period. To indicate the periods where harvesting occurs, we introduce binary variables $d_t \in \{0, 1\}$ and $\delta_t \in \{0, 1\}$ whenever a clearcut or a disaster has occurred. The occurrence of a natural disaster is determined by choosing the rotation period $T$. If the optimal rotation is infinitely long, the regime is CCF, and if it is finite, the regime is RF. The model of Haight and Monserud (1990) is the closest preceding model; however, it specifies the details differently from CCF and the time length between thinnings is constant. The specification in eqs. 1–5 is designed for situations in which the only difference between CCF and RF is that the latter includes clear-cutting and artificial regeneration, while the former does not.

2.2. Stochastic specification with varying initial states

Next we extend the model in eqs. 1–5 by incorporating stochastic stand growth and the risk of sudden natural disasters, such as forest fire, windthrow, and insect outbreaks. These additions lead to several important changes in the model specification compared with the deterministic formulation. First, the state matrix $x_t$ is now stochastic. Second, the ingrowth function $\phi_j$, the diameter increment function $\alpha_{js}$, and the natural mortality function $\mu_{js}$ are also stochastic following the detailed ecological growth model by Pukkala et al. (2013). Although this model has been used in several earlier studies, the stochastic components have not been included in system dynamics with dynamic optimization. Third, we will include the possibility of natural disasters that essentially clear the stand to bare land, leading to artificial regeneration similarly as if the stand has been clear-cut. Including stochasticity requires us to relax the assumption of optimizing an infinite chain of rotations with perfectly equivalent management actions. Closely related to this, we specify the model for any initial stand structure allowing the computation of practically relevant cases in which the initial stand state may be far off the optimal trajectory. Finally, including stochasticity and any initial stand state requires the use of two variables to represent stand state: in addition to $x_t$, which denotes the stand state after harvesting has occurred (as in Section 2.1), we introduce variable $\chi_t$ to represent the stand state at the end of each period after growth but before harvest. This formulation allows a more realistic information on growth stochasticity along with harvesting immediately at the initial state.

In addition, $\bar{x}_{jst}, h_{jst} \geq 0$ for all $j = 1, \ldots, l$, $s = 1, \ldots, n$, $t = t_0, \ldots, T$ and $\bar{x}_{jst,t+1} = 0, j = 1, \ldots, l$, $s = 1, \ldots, n$, $t$ must hold. Equations 2 and 3 represent the development of the mixed-species stand and species interaction arises via the stand density. Equation 4 encodes the fact that $h_{jst} > 0$ for some $j$ and if and only if $\delta_t = 1$.

In this formulation, the optimal choice between RF and CCF is determined by choosing the rotation period $T$. If the optimal rotation is infinitely long, the regime is CCF, and if it is finite, the regime is RF. The model of Haight and Monserud (1990) is the closest preceding model; however, it specifies the details differently from CCF and the time length between thinnings is constant. The specification in eqs. 1–5 is designed for situations in which the only difference between CCF and RF is that the latter includes clear-cutting and artificial regeneration, while the former does not.

Published by Canadian Science Publishing
where the decision variables consist of the number of harvested trees of species \( j \) from size class \( s \), \( h_{j,s,t} \), and the indicator variables \( \delta_{t}^{th} \) and \( \delta_{t}^{cc} \) that determine the timings for thinning and clearcuts, respectively. The stochastic variation in diameter increment and ingrowth are denoted by \( \epsilon_{t} \) and \( \omega_{i} \), respectively. The corresponding functions together with the mortality \( \mu_{j,s}(x_{t}) \) are described in details in the next subsection. The per-period gross profit is defined as

\[
\pi_{t}(x_{t}, h_{t}, \delta_{t}^{th}, \delta_{t}^{cc}) = \begin{cases} 
R_{th}(h_{t}) - C_{th}(h_{t}) - C_{t} & \text{if } \delta_{t}^{th} = 1 \\
R_{cc}(x_{t}) - C_{cc}(x_{t}) - C_{t} - w & \text{if } \delta_{t}^{cc} = 1 \\
-w & \text{if } \delta_{t}^{dis} = 1 \\
0 & \text{otherwise}
\end{cases}
\]

### 2.3. Ecological growth model

Next, we specify the growth functions \( \alpha_{j,s} \), \( \mu_{j,bk} \), and \( \phi_{i} \) for each species \( j = 1, \ldots, l \). Let parameters \( b_{bk,j}, \ldots, b_{bk,7} \) denote the species-specific regression coefficients by Pukkala et al. (2013) reported in Appendix A. First, the mortality function is given by

\[
\mu_{j,s}(x_{t}) = 1 - (1 + \exp(-b_{0j} - b_{1j} \sqrt{d_{s}} - b_{2j}d_{s} - b_{3j} \sqrt{B_{s} \text{pine}(x_{t})}) - b_{4j} \sqrt{B_{s} \text{spruce}(x_{t})} - b_{5j} \sqrt{B_{s} \text{birk}(x_{t}) + B_{s} \text{aspen}(x_{t})}) - b_{6j} \sqrt{B_{s} \text{birk}(x_{t}) + B_{s} \text{aspen}(x_{t}) + B_{s} \text{pine}(x_{t}) - b_{7j}A})^{-1}
\]

where \( B_{s} \) is the basal area for trees with diameters larger than in size class \( s \). Size classes \( s = 1, \ldots, n \) are measured by mean diameter at breast height, \( d_{s} \), from 2.5 to 52.5 cm in 5 cm intervals. Functions \( B_{s,j} = 1, \ldots, l \) represent the basal area in larger Scots pine (Pinus sylvestris L.), Norway spruce (Picea abies (L.) Karst.), silver birch (Betula pendula Roth), and European aspen (Populus tremula L.) size classes.

To specify the fraction \( \alpha_{j,s} \) of trees that move to the next size class during the next 5-year period, we follow Bollandssás et al. (2008) and convert the single-tree-diameter increment models into a transition matrix model by dividing diameter growth with the width \( \sigma \) (5 cm) of the size class \( s \), i.e., \( \alpha_{j,s}(x_{t}) = \sigma \cdot \mu_{j,s}(x_{t}) \).
where \(\omega_{ij,t}\) denotes the residual variation in ingrowth for species \(j\) and 5-year period \(t\). The relative growth variation is especially large among small trees and in the ingrowth estimates. The ingrowths of consecutive 5-year periods and the residuals of predicted ingrowths are positively correlated in the Pukkala et al. (2013) model. The species-specific temporal autocorrelation coefficient is given by \(\hat{p}\). This is largely explained by the fact that one good regeneration year tends to generate ingrowth for several years. In addition to autocorrelation, the residuals of predicted ingrowths are cross-correlated across species, which is captured by the random factors \(u_t\) that are assumed to follow a multivariate normal distribution and be independent within time periods \(t\) with a constant covariance matrix \(\Sigma_u\). Parameter values are based on values presented in Pukkala et al. (2013), and the deterministic ingrowth model is obtained by averaging.

To include the risk of natural disasters, we follow the elementary formulation in Reed (1984). To this end, we have introduced the stochastic indicator variable \(\hat{\delta}_{i}^{\text{act}}\) into the model through \(\text{eq. 10}\). If we have \(\hat{\delta}_{i}^{\text{act}} = 1\), a disaster has occurred during time step \(t\), and all the trees are lost. A regeneration cost occurs at the end of period \(t\). During the \(t_0\) periods, the stand grows and reaches the state specified in the last terms of \(\text{eqs. 8 and 9}\). For simplicity, the stochastic variables \(\hat{\delta}_{i}^{\text{act}}\) are i.i.d. Bernoulli distributed random variables with parameter \(\hat{p}_{\text{dis}}\), which is the probability of a disaster occurring during a period of \(\Delta\) years.

### 3. Computational methods

#### 3.1. Forest management as a RL problem

In RL, the agent (or model) learns by interacting with its environment and by gathering experiences that will help the agent evaluate what was successful and explore what are potentially optimal actions in different situations (Sutton and Barto 2018). The interaction between a learning agent and its environment is defined using the formal framework of Markov decision processes, but in contrast to dynamic programming, its exact mathematical form does not necessarily need to be known. The environment is commonly defined as a large simulation model representing how the actual environment would respond to the actions taken by the agent. In general, RL can be viewed as a special branch of machine learning that is different from both supervised as well as unsupervised machine learning techniques. Although both supervised and RL techniques map input data to outputs, one central difference is that RL methods do not have access to immediate end results or outcomes, but instead they learn from temporary rewards obtained from the environment. When compared with unsupervised learning, RL differs in terms of goals. While unsupervised learning methods work to find similarities and differences between input data points without any access to rewards or correct outputs, the objective in RL is to learn actions that maximize the total discounted cumulative reward of the agent.

In our case, the mathematical representation of the environment simulator is given by the set of constraint \(\text{eqs. 7–11}\) that define the dynamics of the stochastic growth model; see Fig. 1. The agent, on the other hand, may be seen as a decision-maker that makes forest management decisions by following a function \(f_{\pi}: X \rightarrow A\) that maps a given forest stand state to a corresponding optimal harvesting decision, in which the form of the function does not change over time. Here, \(X\) denotes the set of all possible forest stand states and \(A\) is the set of all admissible actions \(a = (h, \hat{\delta}^{\text{th}}, \hat{\delta}^{\text{nc}}) \in A\) that are feasible subject to the model constraints.

As shown in Fig. 1, the agent and the environment interact at discrete time steps \(t = 0, 1, 2, \ldots\). At each time step \(t\), the agent receives a description of the forest stand state \(x_t\), and on that basis selects an action \(a_t = (h_t, \hat{\delta}^{\text{th}}_t, \hat{\delta}^{\text{nc}}_t)\), in which the agent chooses between thinning, clear-cutting, and doing nothing, and how much is thinned. As a consequence of its action, the agent receives a reward, i.e., a per-period profit, \(\pi(x_t, a_t) = \pi(x_t, h_t, \hat{\delta}^{\text{th}}_t, \hat{\delta}^{\text{nc}}_t)\), as defined by \(\text{eq. 12}\), and observes a new stand state \(x_{t+1}\) one time later. The Markov decision process underlying the environment and agent together thereby give rise to a trajectory of states, forest management decisions, and gross profits: \(x_0, a_0, \pi_0, x_1, a_1, \pi_1, \ldots\). In RL, each of these trajectories begins independently of how the previous one ended. As the objective of the agent is to maximize the expected net present value (NPV) over each trajectory, the agent can learn from the rewards it has received by pursuing various forest management policies, as represented by the sequence of actions it has taken. By learning, we refer to the process of how the agent uses trajectory data to update the parameters of its policy function \(f_{\pi}\) that effectively represents a solution to the original stochastic size-structured optimization problem. RL methods may then be understood as mechanisms that specify how the agent’s policy is changed as a result of its experience (cf. Sutton and Barto 2018).

#### 3.2. RL algorithm with parameterized action spaces

The choice of RL algorithms is largely affected by the fact that both action space (the set of all admissible harvesting decisions) and state space (the set of all possible forest stand states) are infinite or continuous. To solve the stochastic optimization problem defined in Section 2.2., we need an algorithm that allows a mixture of continuous (harvesting quantities) and discrete actions (timings of clearcuts and thinnings). We approach the problem of continuous action and state space using the notion of parameterized action spaces suggested by Fan et al. (2019). The idea is to view the overall action as a hierarchical structure instead of a flat set. As shown in Fig. 2, we can decompose each action into two layers. The first layer is a discrete action of choosing between thinning, clear-cutting, and doing nothing. The second layer then chooses the parameters corresponding to each discrete action type coming from the first layer. The parameters represent the actual harvesting quantities that are defined as continuous real-valued variables.

To handle the parameterized action space containing both discrete actions and continuous parameters, Fan et al. (2019) have proposed a hybrid proximal policy optimization (H-PPO) algorithm. The implementation of the algorithm is based on the broadly applied actor–critic framework. To this end, we specify two components: (i) an “actor” function, which the agent uses as its current policy to approximate the unknown optimal policy \(f_{\pi}\), and (ii) a “critic” function, which helps the agent estimate the advantage (benefit) of using the current policy and thereby update the actor’s policy parameters in the direction of performance improvement indicated by the critic. The H-PPO algorithm can be considered an implementation of a stochastic gradient algorithm on the parameter space of stationary policies. Further details on the algorithm are given in Appendix B.

### 4. Results and discussion

#### 4.1. Maximizing the BLV of deterministic stands

Assuming four tree species (spruce, birch, pine, and aspen), regeneration cost \(w = 1401\), a 3% interest rate, and no stochasticity,
we obtain a CCF solution in Fig. 3a. The BLV is €2336.00, which is close to the value €2335 obtained by very different optimization methods in Parkatti and Tahvonen (2020).

A pure pine stand with a 1% interest rate and a €1489 regeneration cost leads to the RF regime depicted in Fig. 3b. The BLV is €18758.97, which agrees with the value €18759 in Parkatti and Tahvonen (2020). While the maximized BLV and the optimal solutions obtained by RL and genetic algorithms and classic optimization methods coincide, the computation times do not. Computing the CCF solution for the four-species case using the latter methods took approx. 170 h in Parkatti and Tahvonen (2020), but only takes approx. 6 min when computed using RL-based methods. One reason behind the difference in computation times is the mixed integer feature of the given optimization problem. Earlier used sequential application of genetic and interior point algorithms to find the integer timing and continuous harvest intensity variables turns out to be slow compared with the very different RL algorithm which solves integer and continuous variables simultaneously.

4.2. Varying the initial stand state

Problem formulations eqs. 6–11 combined with our optimization method make it natural to optimize harvesting from initial stand states beyond bare land. We show that for some initial states, the optimal solution does not fall into pure CCF and RF domains but may include a single clearcut followed by the CCF regime.

Assume a forest with possible initial states (see Table 1) representing a mature spruce stand (initial states A and B). The interest rate is 3% and the regeneration cost €0, implying that the optimal regime is CCF if the initial state was bare land. Given initial state

![Fig. 1. The agent–environment interaction in a size-structured optimization problem.](image1)

![Fig. 2. Parameterized action space.](image2)
A with a rather high number of small diameter trees, the trees in the larger size classes are cut at the first harvest, after which the stand volume converges toward a CCF steady state and no clear-cuts occur (Fig. 4a). In contrast, when the initial state is B (Table 1), with fewer small trees, it is optimal to initially clear-cut but afterwards continue with a CCF regime, as shown in Fig. 4b.

To gain further understanding of how the initial state influences the first harvest, we vary parameters \(a\) and \(b\) to obtain initial state C in Table 1. From Fig. 5, observe that increasing the number of large trees makes clearcutting more favorable. This follows from lower harvesting costs with clear-cutting compared with thinning. Increasing the number of small trees has the opposite effect, as a clearcut removes small trees with high relative value growth and high harvesting costs per cubic metre.

Increasing the regeneration cost decreases the fraction of states where clear-cutting is optimal, as seen by comparing Figs. 5a and 5b.

Given no regeneration cost (Fig. 5a), increasing the interest rate from 1% to 3% increases the profitability of CCF, as the clearcut is followed by a 40-year period with no harvests. This costly revenue postponement dominates the possible higher net revenue from an early clearcut. However, if the interest rate is increased to 5%, and the number of large-diameter trees is high, clear-cutting becomes optimal (\(a = 900, b = 800\) in Fig. 5a) in this case, the lower harvesting cost in the early clearcut dominates the costly 40-year waiting.

---

**Table 1.** Number of trees in pure spruce initial states A, B, and C.

<table>
<thead>
<tr>
<th>Size class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial state A</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Initial state B</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Initial state C</td>
<td>5/12a</td>
<td>5/12a</td>
<td>1/12a</td>
<td>1/12a</td>
<td>0</td>
<td>0</td>
<td>1/2b</td>
<td>1/2b</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Note:** Parameters \(a\) and \(b\) are varied to obtain various initial states C.
Fig. 5. Occurrence of clearcuts with different initial stands. Note: The regeneration costs are (a) \( w = 0 \) for all interest rates, (b) \( w = €1489 \) for \( r = 1\% \), \( w = €1401 \) for \( r = 3\% \). No clearcuts were observed in the case \( r = 5\% \), \( w = €1329 \). [Colour online.]

<table>
<thead>
<tr>
<th>Interest Rate</th>
<th>Number of Small Trees</th>
<th>Number of Large Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>3%</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>5%</td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>

In spite of this, a higher interest rate still favors CCF if the number of large-diameter trees is lower. When the costs of artificial regeneration are included (Fig. 5b), a higher interest rate favors CCF with no exceptions, as the advantage of the free natural regeneration becomes overwhelming.

Earlier size-structured optimization models have shown both numerically and analytically that a higher interest rate favors CCF (Tahvonen 2009, 2015). This result is based on the initial stand as bare land. However, Hyytiäinen and Haight (2010) obtain the same result for an initially stocked stand: a higher discount rate favors CCF. Earlier, in Haigh and Monserud (1990), both a young stand with low volume and old stands with inadequate natural regeneration are immediately clear-cut albeit the long-run regime is CCF. Our results in Figs. 5a and 5b are in line with this finding — increasing the fraction of large trees tends to favor clear-cutting. However, with a higher interest rate and (or) realistic regeneration costs, our results suggest that an initial clearcut may not be optimal for young stands with low volume. Additionally, as the lines for interest rates 3% and 5% cross in Fig. 5a, a high fraction of large trees with a high interest rate favors an initial clearcut. A similar result is obtained in studies without optimizing CCF (Andreassen and Øyen 2002). In our case, the outcome is a consequence of lower harvesting costs in a clearcut. Harvesting costs are not included in Haigh and Monserud (1990) or in Hyytiäinen and Haight (2010).

4.3. Optimizing harvest under stochastic stand growth

Stochasticity in natural regeneration causes major variability of ingrowth. The ingrowth of spruce, along deterministic development with no harvest, converges toward 46 trees (per 5 years). In contrast, the long-term mean ingrowth in stochastic development with no harvest is 62 trees, the standard deviation is 111 trees, and in example realizations the ingrowth varies between 0 and 3080 trees. The ingrowth of aspen follows a similar pattern, while the numbers of birch and pine remain low in spruce-dominated natural stands (without natural disasters). How this natural fluctuation transmits to harvesting decisions and optimal policy has remained an open question.

Assume a four-species stand with a 3% interest rate, regeneration costs of €1401 and bare land as the initial state. Figure 6 shows stochastic realization, in which most of the randomness originates from the stochastic regeneration. It tends to focus on a few favorable years, while most years have low regeneration. High regeneration causes observable changes in species-specific stand volumes. Birch in particular has highly variable regeneration, causing it to temporarily become the dominant species. While the stochastic variation of birch ingrowth is minor in spruce-dominated stands with no harvest (mean 0.8, standard deviation 2), the variation is high in stands with lower densities (mean 251, standard deviation 654), causing high variability in total birch volume, as shown in Fig. 7. Analyzing the solution beyond the initial transient period shows that the mean net revenue per harvest is €3158 with a standard deviation of €557 and that stochasticity decreases the average time between thinnings from 15 to 13 years.

The mean value of the BLV distribution in Fig. 8a equals €2357 and is slightly higher than the optimal value for the deterministic environment (€2336). The distribution is close to a Gaussian distribution. The standard deviation of €51 is over 2% of the expected value, so the natural randomness related to forest growth is enough to create a nontrivial risk. Figure 8b shows the outcome if the policy optimized in the deterministic environment is applied with a stochastic forest. The BLV mean is €2355 and the standard deviation is €49. The minor differences between these distributions imply that the optimal deterministic policy serves as a good approximation of the optimal stochastic policy, i.e., we obtained an outcome close to certainty equivalence albeit the model properties do not satisfy the exact conditions found in the classic paper by Theil (1957). One reason behind our result is that while stochastic variation is high in natural regeneration, the variation is mild in the tree size classes that are actually harvested.

The certainty equivalent result would change after including risk aversion. Several studies (e.g., Clarke and Reed 1989) have applied a nonlinear utility function in the univariate rotation model and have obtained the result that risk aversion shortens optimal rotation. This approach can be viewed as a one tree model since in more general setting adding risk aversion implies optimality to harvest forest in partial cuttings (Tahvonen and Kallio 2006). This suggest that adding risk aversion may enhance the optimality of CCF compared with RF in our size-structured model.
Fig. 6. Sample trajectories of an optimally managed four-species forest with stochastic ingrowth and diameter increment. Note: $r = 3\%$ and $w = 1401$. [Colour online.]

(a)  
(b)  
(c)  
(d)  
(e)  
(f)  

- **spruce**
- **birch**
- **pine**
- **aspen**
Many earlier studies on stochasticity and CCF simultaneously include several sources of stochasticity (Buongiorno and Zhou 2015) or apply very different setups such as the portfolio theory (Paul et al. 2019). These differences make the comparison of their results difficult. For example, it is hard to find any notions on certainty equivalence from earlier stochastic studies. However, according to one main finding in Rollin et al. (2005), stochasticity in stand growth decreases the optimal stand density and increases tree diversity. When we take 200 stochastic realizations, such as those in Fig. 6, stochasticity turns out to increase the average stand volume from 105 to 116 m$^3$·ha$^{-1}$. However, our result for stand diversity supports Rollin et al. (2005): stochasticity increases diversity in a mixed stand with shade-tolerant and -intolerant species.

Our analysis assumes that the current state of the forest is known with certainty. Sloggy et al. (2020) add optimizing costly inventories into the generic univariate rotation model. Following their ideas in mixed species, structured models could be possible.

4.4 Natural disasters

Another type of stochasticity occurs in the form of natural disasters such as storm damage, forest fires, insect outbreaks, and droughts. To analyze their effects on optimal solutions, we assume bare land as the initial state with 1250 spruce and 750 pine trees. The regeneration cost is $w = 0$. In the absence of disasters and with a 1% interest rate, the optimum is the RF solution in Fig. 9a. When the interest rate is increased to 2% or 3%, the regime switches to CCF (Figs. 9b, 9c).
Fig. 9. The optimal solutions without forest fires when (a) $r = 1\%$, (b) $r = 2\%$, and (c) $r = 3\%$. Note: The regeneration costs are 0, and the initial state has 1250 spruce and 750 pine trees in size class 2. [Colour online.]

Fig. 10. The optimal solutions with $r = 1\%$ interest rate and (a) 1%, (b) 2% (pa.) disaster probabilities. Note: The regeneration costs are 0, and the initial state has 1250 spruce and 750 pine trees. [Colour online.]
Next, we set the interest rate to 1% and the annual disaster probability to 1% (or 4.9% per 5 years), which correspond to an average of one disaster per century. The solution generated from the optimal policy in Fig. 10a is very similar to the one with a 2% interest rate and no risk of disasters (Fig. 9b). Similarly, when disaster probability is increased to 2% per year, the optimal solution closely resembles the solution when the interest rate is 3% and no disasters occur (cf. Figs. 9c and 10b).

When the optimal policy is based on the actual stochastic environment (Fig. 11a), the distribution mean is €15 151 and the standard deviation is €3714. As the BLV without disaster occurrence is €21 123, the disaster risk implies a major decrease in BLV. However, notice that in all four cases, it is possible to avoid the disaster over the next centuries and obtain the same BLV, as in the case without any risk of natural disasters (Figs. 11b and 11d), or a somewhat lower BLV, which is the outcome of the policy optimized under risk of natural disasters (Figs. 11a, 11c).

When the initial state is a mature stand instead of bare land, the consequences of disasters may be different. Figure 5 shows that the interest rate effects on optimal CCF/RF choice may depend on the initial stand state, suggesting that the effect of natural disasters may vary as well. In Fig. 12, the initial state corresponds to parameter values $a = b = 600$ in Table 1 and Fig. 5b. Thus, without disasters, an initial clearcut is optimal (Figs. 5b and 12a), but after the inclusion of disasters, the optimal policy leads to CCF (Fig. 12b). Thus, in the cases of mature stands, the probability of...
Fig. 12. Optimal spruce stand development (a) without natural disasters and (b) with a disaster probability of 9.6%. Note: The interest rate $r = 1\%$ and regeneration costs €1489. The initial stand is according to Table 1, where $a = b = 600$. The bare land initial state after clearcut and artificial regeneration has $1750$ trees·ha$^{-1}$ in size class 2. Strategy $a$ has a clearcut at time 0 and strategy $b$ has no clearcuts.

(a)  
(b)

Disasters may still change the optimal policy similarly as increasing the discount rate.

In the widely cited classic paper, Reed (1984) includes the risk of total stand destruction into the generic Faustmann (1849) optimal rotation model. When the probability of destruction is a time-independent Poisson process, the effect is the same as an increase in the interest rate equal to the average rate of destruction occurrence. Within rotation forestry, this result has been extended for models in which the destruction effect depends on stand state, for models with optimized thinning (Reed and Apaloo 1991), and for models with fuel management activities (Amacher et al. 2005). Excluding the latter, the effect of the possible stand destruction is the same, i.e., the rotation shortens. Our result is quite the opposite, and a finite rotation may become infinitely long (i.e., a CCF regime). In our model with the choice between CCF and RF, a higher interest rate favors CCF and the effect of natural stand destructions seems to be similar to an increase in interest rate.

Hyytiäinen and Haight (2010) study the effects of fire on CCF and the choice between the management regimes. According to their main result, the risk of fire increases the relative profitability of converting an initial heterogeneous stand to a single-species plantation (instead of continuing CCF). This result is different from the one we demonstrate in Fig. 12 and is somewhat surprising in light of their other result that a higher interest rate supports CCF. The explanation for this difference in results must be in the model differences and in the Hyytiäinen and Haight (2010) setup specified for wildfires, in which fire damage may be higher in heterogeneous CCF stands. We follow the Reed (1984) generic setup, in which the disaster simply eliminates the existing trees, but note that according to several studies (Hanewinkel et al. 2014), the risk of severe storm damage may be lower in the case of CCF compared with even-aged plantations. We note that including time evolving climate change stochasticity and storm risks would cause a major complication on optimizing forest management.

5. Conclusions

We have applied an optimization approach based on new developments in machine learning complemented with additions needed to handle several specific features of the forest management problem, such as the mixture of both continuous control and binary timing variables. The model can be viewed as an extension of the classic stand-level forest economic problem of maximizing net revenues over an infinite time horizon. Our model design is guided by the aim to increase realism in stand internal structure, mixed species, stochasticity, harvesting cost, harvest timing, and management regimes, but we do not include values of forests beyond timber production nor price stochasticity. The literature on these topics is increasing rapidly but numerous open questions remain. The new optimization method we propose looks promising for further progress in the models’ empirical details and practical relevance without the need to simplify the theoretically sound model structures.

References


Faustmann, M. 1849. Berechnung des wertes, welchen waldboden, sowie noch nicht haubare holzbestände für die waldwirtschaft besitzen [Calculation of the value which forest land and immature stands possess for forestry]. Allgemeine Forst- und Jagdzeitung, 25: 441–455.


Appendix A: Economic and ecological parameter values

Following Parkatti and Tahvonen (2020), trees are divided into 11 size classes s = 1, …, 11, measured by diameter at breast height δt, ranging from 2.5 to 52.5 cm (midpoint) in 5 cm intervals. Each size class has species-specific volumes for both sawtimber vsj and pulpwood vpj, corresponding to the species-specific roadside prices for sawtimber pjl and pulpwood pj2 (Table A1) along with corresponding species-specific roadside prices for sawtimber pjl and pulpwood pj2 (Table A2). Harvesting revenues are then defined separately for sawtimber and pulpwood using species-specific market prices. The gross revenue per period is given by

\[
A1 \quad R(h_t) = \sum_{j} \sum_{v} \left( p_{1j}s + p_{2j}p_{v} \right) h_{tj}
\]

where prices are deterministic.

The harvesting costs depend on species, tree diameter, and the harvested wood quantity. The variable harvesting and hauling costs are derived from the detailed empirical model by Nurminen et al. (2006) for both clear-cutting q = cl and thinning q = th:

\[
A2 \quad C_q h_t = \sum_{j} \sum_{v} \left( c_{qj}s + c_{qj}p_{v} \right) h_{tj}
\]

\[
A3 \quad + c_{q4} \sum_{j} h_{t2j} + c_{q5} \sum_{j} h_{t3j}^{0.7} = q = th, cl
\]

where vl is the total tree volume and cjs are parameters (Table A3). The model defines cutting (eq. A2) and hauling (eq. A3) costs separately. According to this model, the variable harvesting costs increase with total harvested volume but decrease with tree volume. From the parameter values in Table A3, we can see that the cutting costs per tree are higher than clear-cutting, as cjs > cth for all species. Following Parkatti and Tahvonen (2020), we also assume that the smallest size class (with diameter at breast height at 2.5 cm) may only be harvested during a clearcut.

The fixed harvesting cost parameter C1 is set to $500·ha^{-1}$. The fixed cost includes both harvest operation planning cost and the cost of transporting the logging machinery to the site. (present value) cost of artificial regeneration denoted by w is set to $1489·ha^{-1}$ and $1401·ha^{-1}$ for 1% and 3% interest rates, respectively. The regeneration cost parameter is composed of ground mounding at year 0 ($377·ha^{-1}$), planting at year 1 ($739·ha^{-1}$), and...
Can. J. For. Res. Downloaded from cdnsciencepub.com by AALTO UNIV CIVIL & ENVIRONMENTAL on 11/26/21

to represent a suf-

function is not deterministic but a conditional probability density

unknown optimal policy, the function needs to be

actor function. As the actor function is used to approximate the

critic components in more detail. Finally, we present pseudo-code

policy

eqs. 13

Table A1. Size-class specific parameter values for spruce, pine, birch, and aspen according to Heinonen (1994).

<table>
<thead>
<tr>
<th>Size class</th>
<th>$\gamma$</th>
<th>$d$</th>
<th>$v_{1,s,t}$</th>
<th>$v_{2,s,t}$</th>
<th>$v_{1,b,t}$</th>
<th>$v_{2,b,t}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0005</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.0044</td>
<td>7.5</td>
<td>0.0137</td>
<td>0</td>
<td>0.03458</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.0123</td>
<td>12.5</td>
<td>0.06666</td>
<td>0</td>
<td>0.06659</td>
<td>0.07464</td>
</tr>
<tr>
<td>4</td>
<td>0.0241</td>
<td>17.5</td>
<td>0.16990</td>
<td>0</td>
<td>0.1066</td>
<td>0.09764</td>
</tr>
<tr>
<td>5</td>
<td>0.0398</td>
<td>22.5</td>
<td>0.20805</td>
<td>0.23419</td>
<td>0.03905</td>
<td>0.27034</td>
</tr>
<tr>
<td>6</td>
<td>0.0594</td>
<td>27.5</td>
<td>0.30912</td>
<td>0.24475</td>
<td>0.03001</td>
<td>0.48515</td>
</tr>
<tr>
<td>7</td>
<td>0.0830</td>
<td>32.5</td>
<td>0.35975</td>
<td>0.68392</td>
<td>0.02750</td>
<td>0.74205</td>
</tr>
<tr>
<td>8</td>
<td>0.1106</td>
<td>37.5</td>
<td>0.49785</td>
<td>0.96304</td>
<td>0.02647</td>
<td>1.04106</td>
</tr>
<tr>
<td>9</td>
<td>0.1419</td>
<td>42.5</td>
<td>0.50092</td>
<td>1.52313</td>
<td>0.02596</td>
<td>1.38216</td>
</tr>
<tr>
<td>10</td>
<td>0.1772</td>
<td>47.5</td>
<td>0.43424</td>
<td>1.57421</td>
<td>0.02567</td>
<td>1.76537</td>
</tr>
<tr>
<td>11</td>
<td>0.2165</td>
<td>52.5</td>
<td>0.39255</td>
<td>1.89981</td>
<td>0.02549</td>
<td>2.29067</td>
</tr>
</tbody>
</table>

Note: $\gamma$ is the basal area of a tree ($m^2$), $d$ is the mean diameter at breast height (cm), $v_{1,s,t}$ is the pulpwood volume ($m^3$) per tree, and $v_{2,s,t}$ is the sawtimber volume ($m^3$) per tree. Aspen and birch volumes are assumed to be equal.

Table A2. Species-specific roadside sawtimber and pulpwood prices (in $/m^3$).

<table>
<thead>
<tr>
<th>Species</th>
<th>Spruce</th>
<th>Pine</th>
<th>Birch</th>
<th>Aspen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sawtimber</td>
<td>58.44</td>
<td>58.64</td>
<td>49.73</td>
<td>30.16</td>
</tr>
<tr>
<td>Pulpwood</td>
<td>34.07</td>
<td>30.51</td>
<td>30.50</td>
<td>19.74</td>
</tr>
</tbody>
</table>

Note: Prices for spruce, pine, and birch are based on data from the Natural Resources Institute Finland, while the prices for aspen are based on personal communication with the Central Union of Agricultural Producers and Forest Owners.

Table A3. Species-specific parameters for harvesting-cost functions.

<table>
<thead>
<tr>
<th>Species</th>
<th>$q$</th>
<th>$c_{s,a,t}$</th>
<th>$c_{d,a,t}$</th>
<th>$c_{s,b,t}$</th>
<th>$c_{d,b,t}$</th>
<th>$c_{s}$</th>
<th>$c_{d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spruce</td>
<td>th</td>
<td>2.415</td>
<td>0.412</td>
<td>0.758</td>
<td>-0.180</td>
<td>2.272</td>
<td>0.535</td>
</tr>
<tr>
<td></td>
<td>cl</td>
<td>2.100</td>
<td>0.412</td>
<td>0.758</td>
<td>-0.180</td>
<td>1.376</td>
<td>0.393</td>
</tr>
<tr>
<td>Pine</td>
<td>th</td>
<td>2.415</td>
<td>0.547</td>
<td>0.196</td>
<td>0.308</td>
<td>2.272</td>
<td>0.535</td>
</tr>
<tr>
<td></td>
<td>cl</td>
<td>2.100</td>
<td>0.532</td>
<td>0.196</td>
<td>0.308</td>
<td>1.376</td>
<td>0.393</td>
</tr>
<tr>
<td>Birch and aspen</td>
<td>th</td>
<td>2.415</td>
<td>0.420</td>
<td>0.797</td>
<td>0.174</td>
<td>2.272</td>
<td>0.535</td>
</tr>
<tr>
<td></td>
<td>cl</td>
<td>2.100</td>
<td>0.430</td>
<td>0.756</td>
<td>0.174</td>
<td>1.376</td>
<td>0.393</td>
</tr>
</tbody>
</table>

Note: th: thinning; cl: clear-cutting.

tending of the seedling stand at year 11 ($4244-ha^{-1}$) (Parkkati and Tahvonen 2020; Natural Resources Institute Finland 2014, 2018, 2019, 2020).

The parameters of the ecological growth model specified by eqs. 13–16 are given in Table A4.

Appendix B: Proximal policy optimization with parameterized actions

The H-PPO algorithm is based on an actor–critic framework. In this appendix, we will now discuss the definition of actor and critic components in more detail. Finally, we present pseudo-code to describe the key steps in the implementation of the algorithm.

B.1. Actor as approximator for policy function

To formalize this idea, we begin with the specification of the actor function. As the actor function is used to approximate the unknown optimal policy, the function needs to be flexible enough to represent a sufficiently large class of stationary policies. Furthermore, to enable the agent to explore benefits of performing different types of actions, it makes sense to assume that the actor function is not deterministic but a conditional probability density $q_\theta(a|x)$ over the set of all feasible actions $a \in A$ given the current forest stand state $x$. The objective of the H-PPO algorithm is then to find parameters $\theta$ such that the corresponding parameterized policy $\pi_\theta$ generates episodes that maximize the expected NPV, i.e.,

$$
(\text{B1}) \quad \theta^* = \arg\max_{\theta} \mathbb{E}_{\pi_\theta} \left[ \sum_{t=1}^{\infty} \pi(x_t, a_t | \theta) \right]
$$

where the expectation is taken under the assumption that the harvesting actions are chosen using the actor probability distribution $q_\theta$. To use the well-known policy gradient theorem and stochastic gradient ascent to learn the parameters, we further need to assume that the parametric stochastic policies $q_\theta$ are differentiable (Sutton et al. 1999; Marbach and Tissiklis 2001). To ensure that the approximations also have sufficient representational abilities, we have chosen to implement them using neural networks as function approximators, which can be seen as a standard approach in most reinforcement learning applications (see Appendix B.3).
Fig. B1. The neural network structure. The stand structure is given as an input. The actor networks output the discrete harvesting decision and the continuous harvest levels. The value network is used in the advantage estimation. 

\[ V_q(x) = \mathbb{E}_{q_{\theta}} \left[ \sum_{t=0}^{\infty} \pi(x_t, a_t) b^t | x_0 = x \right] \]

is the state value function associated with policy \( q_{\theta} \), which gives the total expected NPV that is encountered while the policy is executed. In an actor-critic algorithm, such as H-PPO, the state value function is also referred to as the critic function, which needs to be approximated by a suitable parametric, differentiable function. The function \( Q_{\theta}(x, a) = \mathbb{E}_{q_{\theta}} \left[ \sum_{t=0}^{\infty} \pi(x_t, a_t) b^t | x_0 = x, a_0 = a \right] \) is the action-value function that assigns to the pair \((x, a)\) the total expected NPV encountered when the decision process is begun in state \( x \), the first action is \( a \), while all the subsequent actions are determined by the policy \( q_{\theta} \). Intuitively, the advantage function can be interpreted as the expected benefit of taking action \( a_0 \) in state \( x_0 \). To find a policy with higher expected NPV, this suggests that policy parameters \( \theta \) should be updated in a direction where they lead us to choose actions with positive advantage values.

In practice, this is done using a generalized advantage estimation (GAE) function proposed by Schulman et al. (2015).

\[ \hat{A}_t^{\text{GAE}(|b|)} = (1 - \lambda) \left( \hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \cdots \right) \]

where the overall advantage is expressed as the sum of 1 to k-step look-ahead functions

\[
\hat{A}_t^{(1)} = -V(x_t) + \pi(x_t, a_t) + bV(x_{t+1}) \\
\hat{A}_t^{(2)} = -V(x_t) + \pi(x_t, a_t) + bV(x_{t+1}) + b^2V(x_{t+2}) \\
\vdots \\
\hat{A}_t^{(k)} = -V(x_t) + \pi(x_t, a_t) + bV(x_{t+1}) + \cdots + b^{k-1}V(x_{t+k-1}) + b^kV(x_{t+k})
\]

where \( \lambda \in [0, 1] \) is a hyper-parameter. When \( \lambda = 1 \), the estimation is known as the Monte Carlo approach. When \( \lambda = 0 \), the definition corresponds to a temporal difference with one step look-ahead.

Collecting the above results together, we can now write the policy gradient as

\[ \nabla_{\theta} \hat{V}_{q_{\theta}} \left[ \sum_{t=0}^{T} \pi(x_t, a_t) b^t \right] \approx \hat{V}_{q_{\theta}} \left[ \sum_{t=0}^{T} \nabla_{\theta} \log q_{\theta}(a_t | x_t) \hat{A}_t^{\text{GAE}(|b|)} \right] \]

where the expectation \( \hat{V}_{q_{\theta}} \) denotes the empirical average over a finite batch of sample trajectories, with \( T \) denoting the maximum length of an observed trajectory. The policy optimization is then carried out by an iterative algorithm that alternates between sampling data from the policy and optimization, which essentially corresponds to a gradient ascent to update the policy parameters. The algorithm is essentially similar to the broadly applied proximal policy optimization (PPO) algorithm with only slight modifications to allow for the use of parameterized actions that combine continuous and discrete decisions (Schulman et al. 2017; Fan et al. 2019).

B.3. Neural network architecture of actor and critic

We use separate actor networks for discrete and continuous actions (see Fig. B1). The two networks are feedforward neural networks that share the first two layers (including the input layer). The forest state variables \( x_{ij, t} \) for all species \( j \) and size classes \( s \) are passed (appropriately normalized) to the input layer. The dimension of the second shared layer is 500. This layer is connected to the separate parts of the two actor networks, which both have input layers of dimension 500 followed by a hidden layer of dimension 200, and an output layer, which is of dimension three (one output for each discrete action) for the discrete actor network, and of the dimension that matches the number of continuous decision variables for the continuous actor network. The activation functions between hidden layers are all of the form \( x \to \max(0, x) \), which is called the rectified linear unit (Goodfellow et al. 2016).

The outputs from the discrete actor network are passed to a softmax-function (Goodfellow et al. 2016) that converts them to discrete action probabilities.

The outputs of the continuous actor network are squeezed to the interval \((0, 1)\) using mapping \( x \to 1/2 \tanh(x) + 1/2 \). These values are used as the expected values of the continuous action distributions for each continuous action. The standard deviations of the Gaussian distributions are directly parameters of the stochastic policy distribution, so there is no separate neural network for them.

The value network used in the critic part of the algorithm is a separate network with hidden layers of sizes 500 and 300. The activation function is the rectified linear unit.
(B5) \[ \theta_{k+1} = \arg \max_{\theta} \mathbb{E}_{q_{\theta_k}}[L(x, a, \theta_k, \theta)] \]

where the surrogate objective \( L \) is defined as the sum of losses corresponding to the discrete and continuous decisions, i.e.,

(B6) \[ L(x, a, \theta_k, \theta) = L_c(x, h, \theta_{k,c}, \theta_c) + L_d(x, \delta^{th}, \delta^{cc}, \theta_{k,d}, \theta_d) \]

where \( a = (h, \delta^{th}, \delta^{cc}), \theta = (\theta_c, \theta_d) \), and

(B7) \[ L_c(x, h, \theta_{k,c}, \theta_c) = \min \left\{ \frac{q_{\theta_c}(h|x)}{q_{\theta_{k,c}}(h|x)}, A^{\text{GAE}(b,a)}(\epsilon) \right\} \]

(B8) \[ L_d(x, \delta^{th}, \delta^{cc}, \theta_{k,d}, \theta_d) = \min \left\{ \frac{q_{\theta_d}(\delta^{th}, \delta^{cc}|x)}{q_{\theta_{k,d}}(\delta^{th}, \delta^{cc}|x)}, A^{\text{GAE}(b,a)}(\epsilon) \right\} \]

(B9) \[ g(\epsilon, A) = \begin{cases} (1 + \epsilon)A & \text{if } A \geq 0 \\ (1 - \epsilon)A & \text{if } A < 0 \end{cases} \]

The clipping done in \( L \) works as a regularizer that penalizes changes to the policy that move the probability ratio away from 1. The hyperparameter \( \epsilon \) corresponds to how far away the new policy can go from the old while still improving the objective. This approach has proven quite effective in ensuring reasonable policy updates. The steps taken in the PPO-Clip algorithm are outlined in pseudo-code as follows:

**procedure** PPO-CLIP

Input: initial policy parameters \( \theta_0 \), initial value function parameters \( \phi_0 \)

for \( k = 0, 1, 2, \ldots \) do

Sample a set of trajectories \( D_k = \{\tau_i\} \) each with \( T \) time steps by running policy \( q_{\theta_k} \) in the environment.

Compute rewards to go \( \hat{U}_t = \sum_{t=0}^{T} \pi(x_t, a_t) b^t \).

Compute advantage estimates \( A^{\text{GAE}(b,a)}(\epsilon) \) based on the current value function \( V_{\phi_k} \).

Update the policy by maximizing the clipped surrogate objective:

\[ \theta_{k+1} = \arg \max_{\theta} \frac{1}{|D_k| T} \sum_{\tau \in D_k} \sum_{t=0}^{T} L(x_t, a_t, \theta_k, \theta) \]

using (stochastic) gradient ascent.

Estimate value function by minimizing mean-squared error:

\[ \phi_{k+1} = \arg \min_{\phi} \frac{1}{|D_k| T} \sum_{\tau \in D_k} \sum_{t=0}^{T} \left[ V_{\phi}(x_t) - \hat{U}_t \right]^2 \]

using gradient descent.

end for

end procedure