

COMPARISON BETWEEN DYNAMIC PROGRAMMING AND REINFORCEMENT LEARNING: A CASE STUDY ON MAIZE IRRIGATION MANAGEMENT

J.-E. BERGEZ

Unité d'Agronomie, INRA, Toulouse, France

E-mail: jberge@toulouse.inra.fr

M. EIGENRAAM

Département of Natural Resources and Environment, Melbourne, Australia

E-mail: Mark.Eigenraam@nre.vic.gov.au

F. GARCIA

Unité de Biométrie et Intelligence Artificielle, INRA, Toulouse, France

E-mail: fgarcia@toulouse.inra.fr

ABSTRACT

Irrigation scheduling is an important decision problem in agriculture. The purpose of the study presented in this paper is to compare dynamic programming (DP) and reinforcement learning (RL) methods for identifying optimal starting irrigation strategies, when a limited amount of water is available for irrigation. Both of the optimization methods use the MODERATO simulator, which includes a growth simulator and an irrigation strategy simulator for maize crops, coupled with a stochastic random weather generator. The results we present illustrate for each of these methods the relations between the number of simulation runs, the size of the discretization and the quality of the approximated optimal strategy that is obtained.

1. INTRODUCTION

Irrigation scheduling in agriculture is an important decision problem that has a major effect on yield, environment and gross margin in water limited areas. Environmentalists and politicians criticise farmers who use a large amount of the available water to irrigate their crops. Applying too much water can potentially create nitrate problems in the groundwater tables. In some irrigation areas where water resources are limited, farmers may receive a fixed amount of water for the growing season or may have to restrict their irrigation flow rates at certain times. For a farmer the profitability of irrigated crops can be improved by reducing the amount of water used and optimizing the timing of application (Stockle and James, 1989). Further, increasing the utilisation of water applied reduces the likelihood of water table accessions, an environmental benefit. New irrigation scheduling approaches, not necessarily based on satisfying the full crop water requirement, but aimed at increasing the efficiency of allocated irrigation water so as to give the highest crop production with the least water use, must be developed (Kirda and Kanber, 1999).

A first attempt to answer these questions is made here. The simple subproblem of deciding when to start the irrigation campaign is formulated. We then compare stochastic dynamic programming and reinforcement learning methods for identifying optimal decision rules. Both of these optimization methods use MODERATO (Bergez et al., 2001), a growth simulator and an irrigation strategy simulator for maize crops, coupled with a stochastic weather generator.

2. MODEL FORMULATION

The problem of optimizing the starting date of the irrigation campaign is a sequential decision problem under uncertainty. Each day, farmers observe the physiological stage of development of the crop and the soil water deficit, a measure of the soil water content. On the basis of these daily observations they must daily decide whether they keep waiting, or start the first irrigation period. Once they start irrigation a strategy is followed that determines an irrigation schedule depending on the weather conditions and on the equipment constraints (flow rate, available water). The crop yields, amount of irrigation water, gross margin and soil dynamics depend stochastically on the weather.

We model the problem of choosing an optimal strategy for starting irrigation as a Markov Decision Problem (MDP). Such a model is defined by a set of possible states, a set of possible decisions, some probabilities describing transitions between successive states given decisions and an objective function to be maximized, defined as a sum of expected returns.

The objective function to be maximized is the expected value of the net revenue after harvest defined as the gross revenue minus irrigation costs. We assume that a limited amount of water is available for irrigation, and that an irrigation strategy has already been specified. The only problem we consider here is to determine an optimal decision rule for starting irrigation.

The state of the process is defined everyday by the two state-variables δ and σ , where δ is the soil water deficit, and σ is the accumulated thermal units above 6°C since sowing. Both variables are continuous. The ranges of δ and σ are respectively the intervals Δ and Σ . As long as there is no irrigation, σ is increasing. When irrigation has started or when σ becomes greater than the upper bound of Σ , the decision process terminates and the final state is s_{end} .

In any state (δ, σ) there are only two possible decisions: wait until the next day (W), and to start irrigation today (S). In the final state s_{end} no decision has to be taken. When the action S is selected, a fixed strategy is applied for the next irrigation decisions. A decision rule for starting irrigation is thus a function π that maps each possible state (δ, σ) in $\Delta \times \Sigma$ to an action W or S .

The net revenue R obtained after harvest is:

$$R = p Y - l N - q C - X, \quad (1)$$

where Y is the final yield, p the price of crop, l the labor cost per irrigation round, N the number of irrigation round, q the unit cost of water, C the total amount of water used for irrigation and X a fixed production cost. Y , N and C depend stochastically on the climate and on the state (δ, σ) where irrigation started.

We denote by $G_S(\delta, \sigma)$ the expected value of the return R obtained when the irrigation is started in state (δ, σ) :

$$G_S(\delta, \sigma) = E[p Y - l N - q C - X]. \quad (2)$$

The daily return of the action W applied in state (δ, σ) is 0, except when this action results in the final state s_{end} . In that case the crop is not irrigated until harvest, and the expected value of the return is

$$G_W(\delta, \sigma) = E[p Y - X]. \quad (3)$$

The Markov property of MDP models requires that the stochastic transition from any state s in $\Delta \times \Sigma$ to s' in $\Delta \times \Sigma \cup \{s_{end}\}$ given the decision d in $\{W, S\}$ is completely determined by the probability $P(s' / s, d)$. For $d = S$ this probability is exactly determined by $P(s_{end} / \delta, \sigma, S) = 1$, $\forall (\delta, \sigma) \in \Delta \times \Sigma$. For $d=W$, the Markov property is an assumption on the climate and soil dynamics.

The nature of this problem is similar to one of optimal stopping (Puterman, 1994). These problems are characterised by a system that evolves uncontrollably (possibly non-stationary) and a decision-maker that only gets to choose the time at which the process terminates. In our case the stopping point is where the decision-maker stops waiting for the next period but commences his/her irrigation schedule. Once irrigation commences no further decisions need to be made and the system evolves as before.

The optimal expected value $V(\delta, \sigma)$ of being in a state (δ, σ) is characterized by the recursive equation

$$V(\delta, \sigma) = \max \{G_S(\delta, \sigma), \int_{\Delta \times \Sigma} V(\delta', \sigma') P(\delta', \sigma' / \delta, \sigma, W) d\delta' d\sigma' + P(s_{end} / \delta, \sigma, W) G_W(\delta, \sigma)\}. \quad (4)$$

The optimal decision rule for starting irrigation is then derived from V as follows:

$$\pi(\delta, \sigma) = S \text{ if } G_S(\delta, \sigma) > \int_{\Delta \times \Sigma} V(\delta', \sigma') P(\delta', \sigma' / \delta, \sigma, W) d\delta' d\sigma' + P(s_{end} / \delta, \sigma, W) G_W(\delta, \sigma), \\ \text{else } \pi(\delta, \sigma) = W. \quad (5)$$

3. OPTIMIZATION PROCEDURES

Two procedures have been considered for deriving optimal decision rules. The first approach relies on a discretization of the domains Δ and Σ of the state variables δ and σ . The discrete transition probabilities are then estimated by simulation, and an approximate numerical solution to equation (4) is obtained by dynamic programming. The second approach, called reinforcement learning does not require an a priori estimation of the transition probabilities. An approximation of the solution to equation (4) is directly obtained by simulation and learning.

Both of the optimization methods use the MODERATO simulator, which is a management oriented cropping system model developed for Maize irrigation (Bergez et al., 2001). MODERATO includes a growth simulator and an irrigation strategy simulator for maize crops, coupled with a stochastic weather generator. MODERATO is able to simulate the soil and crop dynamic, and the stochastic outcomes obtained by following any specified irrigation strategy.

Dynamic programming

In order to use a stochastic dynamic programming algorithm (Kennedy, 1986), the domain $\Delta \times \Sigma$ is represented as a set of grid points (δ_i, σ_j) for $0 \leq i < I$, $0 \leq j < J$. The state-transition probabilities $P(\delta_k, \sigma_l / \delta_i, \sigma_j, W)$ and the expected returns $G_S(\delta_i, \sigma_j)$ and $G_W(\delta_i, \sigma_j)$ can be specified as lookup tables. Note that $P(\delta_k, \sigma_l / \delta_i, \sigma_j, W) > 0 \Rightarrow 0 \leq \sigma_l \leq T_{max} - \delta_i$, and that

$P(s_{end} | \delta_i, \sigma_j, W) > 0 \Rightarrow 0 \leq \sigma_{max} - \sigma_j \leq T_{max} - 6^\circ C$, where T_{max} is the maximum temperature per day and σ_{max} is the upper bound of Σ : many state-transition probabilities are equal to 0.

The values $P(\delta_k, \sigma_l | \delta_i, \sigma_j, W)$, $P(s_{end} | \delta_i, \sigma_j, W)$, $G_W(\delta_i, \sigma_j)$ and $G_S(\delta_i, \sigma_j)$ are estimated by simulation with MODERATO. First a set of trajectories is obtained by always choosing the action W in any state $(\delta, \sigma) \in \Delta \times \Sigma$, until the final state s_{end} is achieved. Then the simulation continues until harvest, without irrigation. A second kind of trajectories are simulated, where the action W is chosen until σ becomes greater than a random threshold σ^l in Σ , and irrigation starts. Then a fixed irrigation strategy is simulated until harvest. The visited points (δ_t, σ_t) on these trajectories and the observed revenues R_t are collected and used for calculating the maximum-likelihood estimates of the state-transition probabilities $P(\delta_k, \sigma_l | \delta_i, \sigma_j, W)$ and $P(s_{end} | \delta_i, \sigma_j, W)$, and the expected returns $G_S(\delta_i, \sigma_j)$ and $G_W(\delta_i, \sigma_j)$ for $P(s_{end} | \delta_i, \sigma_j, W) > 0$.

This optimal stopping problem with IJ states and 2 actions can be directly solved using the policy iteration algorithm in infinite horizon, with no discount factor¹.

Reinforcement Learning

The reinforcement learning approach consists in modifying iteratively an estimate of the optimal value function on the basis of simulated state-transitions and returns (Sutton and Barto, 1998). For the optimal stopping problem we consider, the principle of Q-learning, the most studied reinforcement learning algorithm, is to estimate the Q-values

$$Q(\delta, \sigma, W) = \int_{\Delta \times \Sigma} V(\delta', \sigma') P(\delta', \sigma' | \delta, \sigma, W) d\delta' d\sigma' + P(s_{end} | \delta, \sigma, W) G_W(\delta, \sigma),$$

$$Q(\delta, \sigma, S) = G_S(\delta, \sigma). \quad (6)$$

At the beginning of each simulation, like for dynamic programming, a random threshold σ^S is chosen and the action W is followed as long as $\sigma < \sigma^S$. When $\sigma \geq \sigma^S$ in the current state (δ, σ) , the action S is applied and a fixed irrigation strategy is simulated. At the end of such a trajectory $\{(\delta_0, \sigma_0), (\delta_1, \sigma_1), \dots, (\delta_t, \sigma_t)\}$ a return R_t is received and the Q-values are updated with the following update rule:

$$Q(\delta_k, \sigma_k, W) \leftarrow Q(\delta_k, \sigma_k, W) + \varepsilon (\max\{Q(\delta_{k+1}, \sigma_{k+1}, W), Q(\delta_{k+1}, \sigma_{k+1}, S)\} - Q(\delta_k, \sigma_k, W)), k < t$$

$$Q(\delta_t, \sigma_t, d_t) \leftarrow Q(\delta_t, \sigma_t, d_t) + \varepsilon (R_t - Q(\delta_t, \sigma_t, d_t)). \quad (7)$$

In this update rule, the last decision d_t is equal to W or S depending on the value of σ^l . The stepsize ε decays towards 0 with the number of simulations.

In order to apply this algorithm, the domain $\Delta \times \Sigma$ is represented as p shifted sets of grid points (δ^k_i, σ^k_j) for $0 \leq i < I$, $0 \leq j < J$, and $1 \leq k \leq p$. The Q-value functions are then approximated by:

$$Q(\delta, \sigma, d) = \sum_{k=1}^p Q(\delta^k_i, \sigma^k_j, d) \text{ for } d \in \{W, S\}, \quad (8)$$

where (δ^k_i, σ^k_j) is the closest grid point to (δ, σ) on the grid k . In that case the Q-learning update rule (7) is adapted and turns now on the $Q(\delta^k_i, \sigma^k_j, d)$ values.

¹ General Purpose Dynamic Programming (GPD, Kennedy, J) software is used to find the optimal solution. This software and supporting documentation are available at the following URL. <http://www.cornerstone-computing.com/gpd/gpd.html>.

4. NUMERICAL APPLICATION

To test the specified methods, we used the simulator MODERATO on a specific case based on data from southwestern France. The soil is a medium clay-silt soil with clay accumulation at 0.8 m depth and a 150 mm available water capacity, locally called « Boulbènes moyennes ». This type of soil is representative of a large area of the Midi-Pyrénées. The climate data used by the weather generator (Racsco et al., 1991) is a 49-year daily weather record from Blagnac near Toulouse. Sowing is between 15 April and 15 May as soon as the cumulative rainfall during the previous 3 days is less than 15 mm. Variety Volga is sown at 80 000 plants/ha. The irrigation equipment allows a 5 mm/day flow rate and a limited 200 mm total amount of water is available for irrigation. When irrigation had started, the other rounds are based on a 7-day frequency and on a 30 mm depth per irrigation. If rainfall occurs during the irrigation campaign, the round is delayed by 1 day every 5 mm of rainfall. After 01 September, if the soil water deficit is less than 50 mm, the irrigation campaign stops, otherwise a last irrigation round is performed. The crop is harvested when grain moisture content reaches 20% or cumulative thermal units from sowing reach 2100 °C day and if the cumulative rainfall during the previous 3 days is less than 15 mm. In any case, the crop must be harvested before 15 October.

The average selling price for maize is assumed to be 700 F/tonne in the Toulouse area. Fixed costs (seeds, weeding, fertilizer, insurance) are assumed to be 2150 F/ha. The cost of irrigation water is assumed to be 5 F/mm and that of setting up a new irrigation round is assumed to be 50 F (labor). The state variables δ and σ range respectively from 0 to 150 mm and from 0 to 1800°C day.

We used the DP and RL methods with several values of the number of grid points I, J , and of the total number N of MODERATO simulated years for estimating the DP model, or for learning the optimal value function. Note that the extra CPU time used by the policy iteration algorithm is equivalent to the time needed for simulating about 100 years (few seconds).

Figure 1 represents the average value of the best policies obtained by dynamic programming and reinforcement learning with several values of the parameters I, J and N , obtained by simulating them on 1000 years. A first conclusion seems to be that RL performs better than DP when a small number of simulations is available. When N is sufficiently large ($N > 100000$), all the policies are equivalent. Furthermore, better results are obtained for small I, J values.

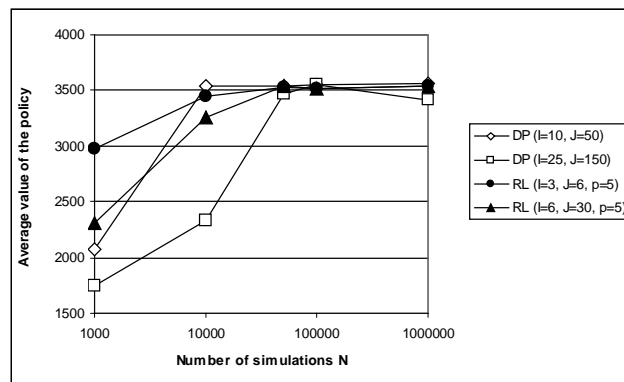


Figure 1: Average value of the DP and RL policies, at several values of N, I, J .

Figure 2 shows 1000 simulated starting points for irrigation obtained by using the 3×6×5 RL policy and the 10×50 DP policy after $N=10^6$ years. We note that most of these points all roughly located between 700 and 900°C day.

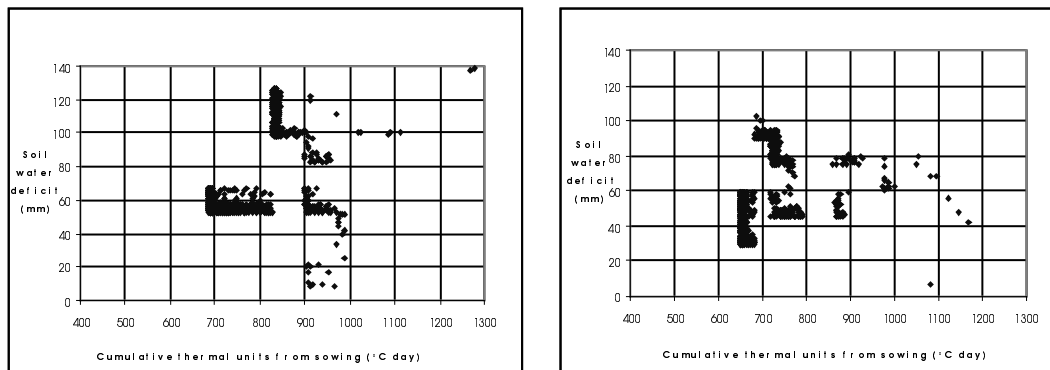


Figure 2: Simulated starts of irrigation with the 3×6×5 RL policy and the 10×50 DP policy, $N=10^6$ simulations

5. CONCLUSION

The kind of policy obtained by this preliminary study (starting irrigation by 700 °C day) is consistent with the expert knowledge of irrigation practitioners. Two methodological results can be underlined: (i) RL performs better than DP when only a small number of simulation runs are available ; (ii) for a given number of simulation runs, a smaller number of grid points is preferable. We now intend to apply this approach to characterize the optimal policies for different irrigation constraints (total amount of water for irrigation, soil available water capacity) and to extend the approach to the whole irrigation strategy (Bergez et al., 2001). Further, the opportunity cost to the irrigator of limiting water applications for environmental reasons can be explored using the techniques we have described above.

6. REFERENCES

- Bergez, J.-E., Debaeke, Ph., Deumier, J.-M., Lacroix, B., Leenhardt, D., Leroy, P., Wallach, D., 2001. MODERATO: an object-oriented decision model to help on irrigation scheduling for corn crop. *Ecological Modelling*, 137(1): p43-60.
- Kennedy, J.O, 1986. *Dynamic Programming: application to agricultural and natural resources*, Elsevier Applied Science, London.
- Kirda, C., Kanber, S., 1999. Water, no longer a plentiful resource, should be used sparingly in irrigated agriculture. In *Crop yield responses to deficit irrigation*, Kirda, Moutonnet, Hera and Nielsen Eds, Kluwer academic publishers, p1-20.
- Puterman, M.L. 1994. *Markov Decision Processes*, Wiley, New-York.
- Racsko, P., Szeidl, L. and Semenov, M., 1991. A serial approach to local stochastic weather models. *Ecological Modelling*, 57.
- Stockle, C.O., James, L.G., 1989. Analysis of deficit irrigation strategies for corn using crop growth simulation. *Irrigation Science*, 10: p85-98.
- Sutton, R.S., Barto, A.G., 1998. *Reinforcement Learning: an introduction*, MIT Press, Cambridge.