13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13
Seoul, South Korea, May 26-30, 2019

Direct policy search as an alternative to POMDP for sequential decision problems in infrastructure planning

Elizabeth Bismut
PhD Student, Engineering Risk Analysis Group, Technische Universität München, Germany

Daniel Straub
Professor, Engineering Risk Analysis Group, Technische Universität München, Germany

ABSTRACT: Most infrastructure planning challenges belong to the class of sequential decision problems, characterized by significant initial uncertainty on the demand and performance of the system, and the possibility to collect information and reduce the uncertainty throughout the service life. In this paper, we consider a generic infrastructure planning problem and compare two main solution frameworks, partially observable Markov decision processes (POMDPs) and direct policy search (DPS) with a choice of heuristics. A case study is set up so that the the belief space is described by only two parameters and the POMDP approach yields an exact solution. We investigate the performance of direct policy search by examining the optimal choice of the heuristics through a comparison with the POMDP solution. Depending on the type of system and reward function considered, parameters defining the heuristics can be thresholds on the demand or the system reliability, after which intervention is required, or critical damage values that suggest a component repair. The choice of the parameters directly influences the goodness of the solution found. We identify key factors for the optimal selection of the heuristic parameters for the generic problem, and provide insights into which specific features of the system that guide the decision process.

1. INTRODUCTION

Good infrastructure planning ensures a continued operation and the safety of its users at a minimal cost. Planning problems typically are associated with uncertainty, be it on population fluctuation for the planning of a waste water system or on annual maximum precipitations in a floodplain for designing flood protection measures. Furthermore, the costs associated with demands exceeding capacity can be very steep. Regular monitoring and data collection can reduce the uncertainty and improve the predictions of future demand. These planning problems belong to the class of sequential decision problems, and are computationally challenging to solve. The optimal strategy should consider at every time step all relevant past information about the system when deciding on an action, as well as the potential information gain at future time steps.

The general decision problem under uncertainty was formalized with Bayesian decision theory (Raiffa and Schlaifer, 1961). Two main solution frameworks have been developed. The first, partially observable Markov decision process (POMDP), attempts to provide a universal plan and, in theory, accounts for all possible scenarios (Howard, 1960; Kaelbling et al., 1998). POMDP modeling has been adopted for simple demand/capacity problems (Špačková and Straub, 2017; Pozzi et al., 2017), but in practice becomes intractable once the scale of the problem increases (Papadimitriou and Tsitsiklis, 1987). The second framework is direct policy search (DPS) with a choice of heuristics, and makes the optimization tractable by finding a strategy that is globally suboptimal, but is optimal within a reduced solution space. DPS is a well-known alternative in the field of robotics (Shani et al., 2013; Pooya and Pakdaman, 2018), and is often chosen for its flexibility and intuitive principles, notably for risk based inspection (RBI) planning (Straub, 2004; Kim and Frangopol, 2010; Bismut and Straub, 2018; Luque and Straub, 2019).

DPS operates in a different state space than POMDP. Its performance depends on the choice of heuristics, but also on the system and cost model parameters. In this paper, we define a generic decision problem and compare the performance of three different DPS heuristics with the optimal strategy computed with POMDP. In particular, we identify the factors which influence the goodness of the DPS solutions, and establish a comparison, not only at the cost level, but also at the policy level between the exact POMDP solution and the DPS approach.

2. A GENERIC PLANNING PROBLEM

2.1. The case study

We consider the following generic infrastructure planning problem, adapted from Straub and Špačková (2016). In every year t, the system capacity a_t should to cover the demand θ_t, which is expected to increase over time. The initial system capacity a_1 is fixed by the decision-maker, and can be increased at any time step, at a cost. Failure to meet the demand (within a certain margin) incurs a penalty. A strategy must give the decision-maker an answer to the following questions at every time step: "Should the capacity be increased? If yes, by how much?" The optimal strategy maximizes the expected life-cycle utility.

Examples that fall within this class of problems include waste water treatment plants and flood protection subject to uncertainty in population evolution and climate change, or asset management of structural systems subject to uncertain deterioration and loads.

The decision process is formally illustrated by the influence diagram in Figure 1. Circular nodes represent random variables, square nodes are decisions on the system capacity and diamond-shaped nodes are the initial capacity cost and following capacity upgrading costs.

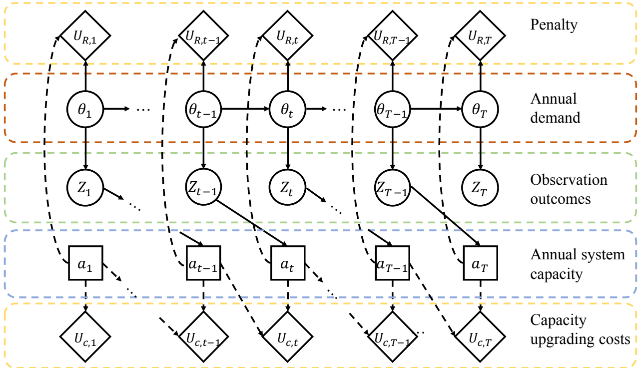

Figure 1: Influence diagram for the planning problem.

After fixing the initial capacity a_1, the system is subject to the first year demand θ_1; Z_1 is the noisy observation of the demand θ_1; the capacity a_2 for next year is then decided, and so on up to the time horizon $T = 100$ [years].

The process is here Markovian, as the state of the system at time t is independent of the past states given the state of the system at time $t-1$.

The available capacities are restricted to the values $\{0,1,2,3,4,5,6\}$. Here, the capacity cannot decrease over time, i.e. $a_t \geq a_{t-1}$.

Table 1 details the model assumptions for the initial distributions and the state transitions over time.

Table 1: Model parameters.

Variable	Type	Mean	Std. Dev.
θ_1	Normal dist.	$\mu_{ini} = 1.0$	$\sigma_{ini} = 1.0$
θ_t	Function	$\theta_{t-1} + \tau$	-
τ	Normal dist.	$\mu_\tau = 0.02$	$\sigma_\tau = 0.05$
Z_t	Normal dist.	θ_t	$\sigma_\varepsilon = 0.1$
T	Deterministic	100 [years]	-

The assumption of a Gaussian distribution for the demand is not very realistic as it gives a non-zero probability for the demand to be negative. Nevertheless, for the purpose of this study, and to allow exact computation of the solution to this decision problem, the normal distribution is not truncated at 0.

The cost model is summarized in Table 2. It can also be adapted to consider varying degrees of system flexibility (Straub and Špačková, 2016). The parameters of Table 2 correspond to full flexibility.

Table 2: Cost model.

Initial capacity cost	$U_{C,1} = -c_a \cdot a_1$
Capacity upgrading for $t \geq 2$	$U_{C,t} = -(a_t - a_{t-1}) \cdot c_a \cdot \gamma_t$
Exceedance penalty	$U_{R,t} = -\Phi(\frac{a_t - \theta_t}{\alpha}) \cdot c_F \cdot \gamma_{t-1}$
Upgrading cost factor	$c_a = 1$
Penalty factor	$c_F = 10$
Discount factor	$\gamma_t = \frac{1}{(1+r)^t}$ with $r = 0.02$
Tolerance	$\alpha = 0.1$

Φ denotes the standard normal cumulative distribution function.

2.2. *The optimization problem*

The optimal strategy for this decision problem is

$$\mathcal{S}^* = \arg\max_{\mathcal{S} \in \mathscr{S}} (\mathbf{E}[U(\mathcal{S})]), \tag{1}$$

where $\mathscr{S}$ is the space of all possible strategies, and $\mathbf{E}[U(\mathcal{S})]$ is the expected total life-cycle utility associated with strategy $\mathcal{S}$. The two approaches to solving this equation, or to approximating its solution, are described in the following sections.

3. POMDP

3.1. *Theoretical solution*

The POMDP approach divides the optimization problem in Eq. 1 into sub-problems and solves it recursively by dynamic programming and backwards induction. This was first proposed for Markov decision processes (MDPs) by Bellman (1957), who considered the problem of inspection and maintenance planning involving a production machine subject to defects and failures.

POMDPs only provide partial information about the state of the system at each time steps, however it can be shown that POMDPs are MDPs over the belief B_t, which is defined as the probability distribution over the state of the system given all the past information collected and all previous actions (Kochenderfer, 2015). Essentially, it condensates all the knowledge about the system into one variable, the belief, which is then fully observable at every time step. The equivalent MDP is illustrated in Figure 2.

Approximate solutions for POMDPs are available (Ross et al., 2008; Erez and Smart, 2012; Papakonstantinou et al., 2018), but their performance decreases with increasing dimensionality of the problem.

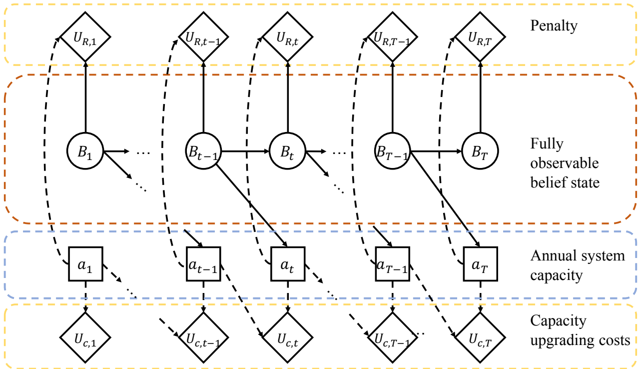

Figure 2: Influence diagram for the planning problem translated into the belief space.

3.2. *Exact solution for Gaussian POMDPs*

In the case study defined in Section 2.1, the POMDP problem is defined as a Gaussian POMDP, as all the probability distributions are assumed to be Gaussian. This Gaussian POMDP case study was developed in (Liu, 2017) and further investigated by Vahidi (2018). The belief B_t, under imperfect and partial observation, can be exactly described by the mean μ_t and standard deviation σ_t of $\theta_t | Z_{1:t}$, and by the current capacity a_t. Eq. 2 and Eq. 3 define the belief state transition. Notably, the standard deviations σ_t are deterministically defined in a recursive manner, hence the belief B_t consists only of $\{\mu_t, a_t\}$.

$$\begin{cases} \mu_{t+1} & \sim \mathrm{N}\left(\mu_t + \mu_\tau, \frac{\sigma_t^2 + \sigma_\tau^2}{\sqrt{\sigma_t^2 + \sigma_\tau^2 + \sigma_\varepsilon^2}}\right) \\ \mu_1 & \sim \mathrm{N}\left(\mu_{ini}, \frac{\sigma_{ini}^2}{\sqrt{\sigma_{ini}^2 + \sigma_\varepsilon^2}}\right) \end{cases} \tag{2}$$

and

$$\begin{cases} \sigma_{t+1} & = \sqrt{\frac{\sigma_\varepsilon^2 \cdot (\sigma_t^2 + \sigma_\tau^2)}{\sigma_t^2 + \sigma_\tau^2 + \sigma_\varepsilon^2}} \\ \sigma_1 & = \frac{\sigma_{ini} \cdot \sigma_\varepsilon}{\sqrt{\sigma_{ini}^2 + \sigma_\varepsilon^2}}. \end{cases} \tag{3}$$

The variable μ_t is discretized on $[-\infty + \infty]$, with 299 intervals between 0 and 6. The conditional probability tables for every time step $1 \leq t \leq T$ are computed based on Eq. 2 and Eq. 3. The reward function R is defined by the cost model in Table 2. The equivalent MDP is solved by value iteration (Bellman, 1957; Kochenderfer, 2015).

The solution of the POMDP consists of policy tables for every time step $1 \leq t \leq T$ which associate a capacity to every belief state of the system. One of these tables is illustrated in Figure 3.

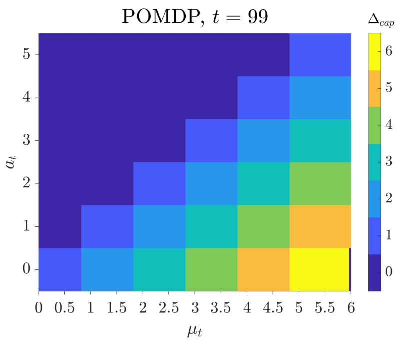

Figure 3: Policy table in the belief space $\{\mu_t, a_t\}$. The increase in capacity between a_t and a_{t+1}, Δ_{cap}, is represented by a corresponding color.

4. DIRECT POLICY SEARCH

4.1. Methodology

Direct policy search (DPS) consists in restricting the space of strategy $\mathscr{S}$ in Eq. 1, by choosing a heuristic $\mathcal{W}$ to explore only a subspace of $\mathscr{S}$ (Sutton et al., 2000). In this approach, heuristic parameters $\boldsymbol{w} = \{w_1, w_2, \dots w_l\} \in \mathcal{W}$ specify the strategy $\mathcal{S}_{\boldsymbol{w}}$, which deterministically assigns an action a_t at every time step t, given the past observation outcomes and past actions. Example of heuristics are given in Section 4.3. The solution $\mathcal{S}^*$ to the optimization problem in Eq. 1 is therefore approximated by $\mathcal{S}_{\boldsymbol{w}^*}$, where

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w} \in \mathcal{W}} (\mathbf{E}[U(\mathcal{S}_{\boldsymbol{w}})]). \tag{4}$$

Compared to the POMDP approach, DPS has the added advantages that it does not require the process to be Markovian, and tends to perform better as the scale of the problem increases. Furthermore, the intuitive formulation of the strategy with simple heuristics makes DPS a privileged solution in decision planning, as opposed to the "black box" approach of the POMDP solution.

4.2. Global optimization method for DPS

The DPS optimization in Eq. 4 requires a global optimization method. For this case study, we propose to use the cross entropy (CE) method to find the heuristic parameters that maximize the expected total life-cycle utility. This method is well adapted to objective functions that are computationally challenging to evaluate exactly, such as expected values. CE can handle noisy objective functions and still converge towards the point which minimizes the original objective function (Kochenderfer, 2015).

4.3. Heuristics investigated

We investigate three heuristics to compare to the exact solution, which are commonly adopted in sequential decision making. As in the POMDP setup, the capacity can only increase over time, and must take its values in $\{0,1,2,3,4,5,6\}$. The following algorithms detailing the heuristics do not include this last constraint for clarity.

After deciding on the initial capacity a_1, Heuristic 1 simply increases the capacity at fixed time intervals ΔT by a constant Δa, without integrating any knowledge from the noisy observations of the demand. For this heuristic, the CE method optimizes the values of a_1, ΔT and Δa.

Heuristic 1: Upgrade the capacity at fixed time intervals

in : heuristic parameters $\{a_1 \geq 0, \Delta a \geq 0, \Delta T\}$
time horizon T
out: vector a_t of capacities
$t \leftarrow 2$;
while $t \leq T$ **do**
  if $t \mod \Delta T = 0$ **then**
    $a_t \leftarrow \min(a_{t-1} + \Delta a, 6)$;
  end
  $t \leftarrow t + 1$;
end
return a_t

For Heuristic 2, the initial capacity a_1 is set. At every time step t, an observation of the demand Z_t is acquired, and the probability that the demand at the next time step, θ_{t+1}, exceeds the current system capacity, a_t, is computed conditional on the observations up to that time, $Z_1..Z_t$. If this probability is bigger than a threshold p_{th}, the system capacity is increased by at least Δa. In the algorithm, $F^{-1}_{\theta_{t+1}|Z_1..Z_t}$ is the inverse cumulative distribution function of θ_{t+1} conditional on $Z_1..Z_t$. The CE method optimizes the values of a_1, Δa and p_{th}.

Heuristic 2: Upgrade the capacity based on a probability of exceeding demand

```
in  : heuristic parameters {a_1 ≥ 0, Δa ≥ 0, p_th}
      time horizon T, observations of
      demand Z_1...Z_T
out: vector a_t of capacities
t ← 1;
while t < T do
    if Pr[θ_{t+1} > a_t|Z_1..Z_t] > p_th then
        a_th ← F^{-1}_{θ_{t+1}|Z_1..Z_t}(1 − p_th);
        a_{t+1} ← min(max(a_t + Δa, a_th), 6);
    end
    t ← t + 1;
end
return a_t
```

Heuristic 3 differs from Heuristic 2 in the condition for capacity upgrading. The system capacity is increased by at least Δa if the current observation Z_t is within a certain margin of the available capacity a_t. This margin is defined by factor k. The CE method optimizes the parameters a_1, Δa and k.

Heuristic 3: Upgrade the capacity based on the observed value of demand

```
in  : heuristic parameters {a_1 ≥ 0, Δa ≥ 0, k}
      time horizon T, observations of
      demand Z_1...Z_T
out: vector a_t of capacities
t ← 1;
while t < T do
    if a_t − Z_t < k · a_t then
        a_th ← k · a_t + Z_t;
        a_{t+1} ← min(max(a_t + Δa, a_th), 6);
    end
    t ← t + 1;
end
return a_t
```

5. COMPARISON OF RESULTS

5.1. *Optimal strategy with DPS and POMDP*

The optimal parameter values for the different heuristics are found with the CE optimization method:

- for Heuristic 1, $a_0 = 3$, $\Delta a = 1$, $\Delta T = 65$ [years];
- for Heuristic 2, $a_0 = 2$, $\Delta a = 1$, $p_{th} = 1.16 \cdot 10^{-5}$;
- and for Heuristic 3, $a_0 = 2$, $\Delta a = 1$, $k = 0.1236$.

Figure 4 shows how these strategies and the POMDP solution strategy react to the same observation history. This figure also depicts the mean estimate of the demand conditional on the full observation history.

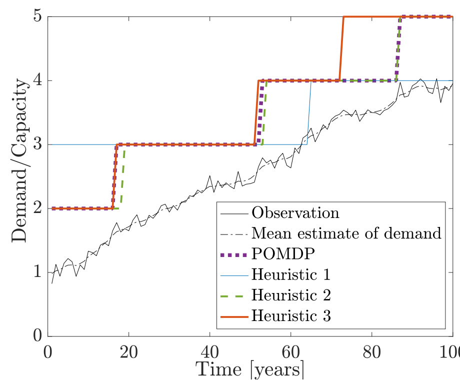

Figure 4: Sample observation history, associated mean estimate of the demand, and prescribed capacity according to the POMDP solution and the three heuristic solutions.

5.2. *Comparison of expected cost*

The expected costs are summarized in Table 3 and illustrated in the bar chart in Figure 5. As expected,

Table 3: Expected costs.

Solution	Exp. cost of upgrading	Exp. penalty	Total exp. cost
Heur. 1	3.28	$9.7 \cdot 10^{-1}$	4.25
Heur. 2	2.72	$3.3 \cdot 10^{-2}$	2.75
Heur. 3	2.76	$2.9 \cdot 10^{-2}$	2.79
POMDP solution	2.71	$3.5 \cdot 10^{-2}$	2.75

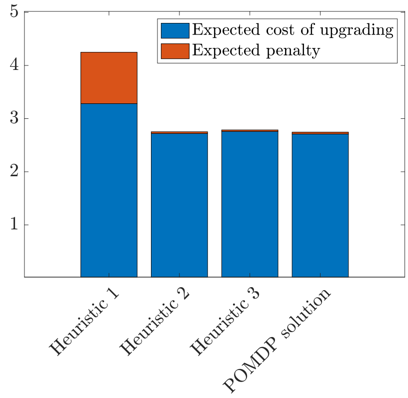

Figure 5: Expected total life-cycle utility for the investigated heuristics and the reference POMDP solution. The cost is split between the expected penalty and the expected capacity upgrading cost.

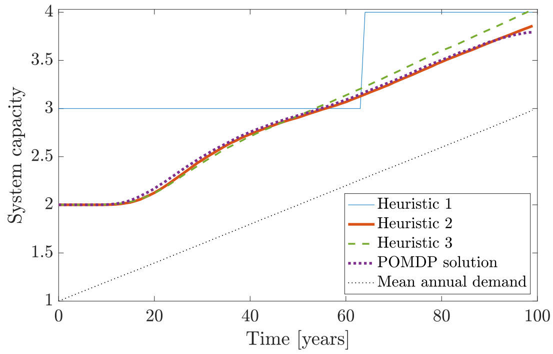

Figure 6: Evolution of the mean optimal capacity throughout the service life for the investigated heuristics and the reference POMDP solution.

Heuristic 1 does not perform as well as the other two, as it does not include any information about the demand throughout the lifetime. Heuristic 2 achieves the best expected cost among the heuristics investigated, with a total expected cost virtually equal to expected cost of the POMDP solution. Heuristic 3 does not integrate all the knowledge from past observation and the associated uncertainty, however it approximates the POMDP solution quite well. This is due to the low uncertainty on the observation, and it is expected that the performance of this heuristic strongly depends on that uncertainty.

5.3. *Comparison of expected capacities*

Figure 6 compares the evolution of the mean optimal capacities for the DPS approach and the POMDP solution. The optimal strategies for Heuristics 2 and 3 and the POMDP solution start with the same initial capacity $a_1 = 2$, and follow the same trend. Heuristic 3 deviates from the exact solution around time step $t = 55$, and Heuristic 2 follows very closely the evolution of the POMDP solution. This is consistent with the comparison of the expected costs.

5.4. *Comparison of policies*

The yearly policies and the corresponding policy tables of the optimal heuristics can be obtained for each time step from their definitions in Section 4.3.

For Heuristic 1, the policy translated into the belief space $\{\mu_t, a_t\}$ depends only on a_t and time step t. The policy tables are trivial, as they prescribe a unit increase in capacity for year 65 and no increase in capacity for the other years.

The conditional clause in Heuristic 2 is directly translated into a condition on the observed value of the belief $\{\mu_t, a_t\}$, as

$$\Pr[\theta_{t+1} > a_t | Z_1..Z_t] = 1 - \Phi\left(\frac{a_t - \mu_t - \mu_\tau}{\sqrt{\sigma_t^2 + \sigma_\tau^2}}\right). \quad (5)$$

Hence the policy following Heuristic 2 is deterministic for any belief state $\{\mu_t, a_t\}$.

Strategies described by Heuristic 3 do not translate well into the belief space representation. Indeed, contrary to Heuristics 1 and 2 which attribute to every belief state $\{\mu_t, a_t\}$ a determined capacity, the conditional clause $a_t - Z_t < k \cdot a_t$ implies a stochastic policy in the belief space. The seemingly easier rule of acting on an observation gives a deterministic action in the state space, but only a probabilistic action in the belief space (it actually depends on the belief state at the previous step). Heuristic 3 is not futher considered in this section.

Figure 7 compares the policy tables at the beginning and end of service life for Heuristic 2 and the POMDP solution. While their policies are visibly not equivalent over the full belief space, they agree quite well on the limits of the zone defined by $\Delta_{cap} = 0$, especially for $t = 1$.

Heuristic 2 is almost stationary in the belief space. This can be explained by examining Eq. 5, and noticing that σ_t is the only time-varying parameter of the policy. Eq. 3 shows that σ_t converges after a few time steps to a constant value.

By comparing the POMDP policies at time steps $t = 1$ and $t = 99$, we notice that the zone delimiting $\Delta_{cap} = 0$ shifts towards higher values of μ_t. This is explained by the discount factor which gives less importance to rewards at the end of the service life, hence the solution tends to accept a higher risk of demand exceeding capacity at the end of service life. However the cost function parameters are such that the variation in the policy tables is significant only after time step $t = 95$ [years], making the POMDP policy also quasi-stationary. By varying the ratio between the penalty factor c_F and the upgrading cost factor c_a, the exact solution might present a bigger variability in the yearly policies.

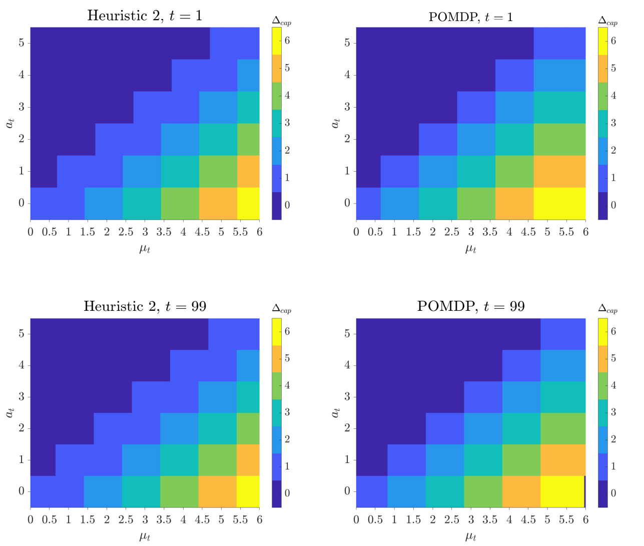

Figure 7: Policy $a_{t+1}(\mu_t, a_t) = a_t + \Delta_{cap}$, for the optimal strategy of Heuristic 2 (left) and the POMDP solution (right), at time steps $t = 1$ (top) and $t = 99$ (bottom).

6. DISCUSSION

In this case study, Heuristics 2 and 3, both with three parameters, give very good approximations of the exact solution found with the POMDP approach.

Further studies are required to understand the roles of the various characteristics of the model towards the goodness of the approximation. For example, the uncertainty on the observation of demand is small, with a coefficient of variation of 10% at the beginning of life, reaching 3% at the end of life. It is expected that the performance of Heuristic 3 deteriorates in comparison to Heuristic 2 as the uncertainty in the observation increases. Furthermore, as the case study in this paper leads to quasi-stationary policies, more complex cost models should also be considered. More crucially, the case study was based on a purely linear demand model, which might be a reason for the good performance of heuristic strategies defined by few parameters.

7. CONCLUSIONS

We compared three different heuristics with the direct policy search approach to solve a generic sequential decision problem for which the exact solution was calculated using the POMDP approach. We found that the heuristics yielded a close approximation of the exact solution, and identified different factors in the model that influence the goodness of the solution.

REFERENCES

Bellman, R. (1957). "A Markovian decision process." *Journal of Mathematics and Mechanics*, 679–684.

Bismut, E. and Straub, D. (2018). "Adaptive direct policy search for inspection and maintenance planning in structural systems." *Sixth International symposium on Life-Cycle Civil Engineering 2018, IALCCE.*

Erez, T. and Smart, W. D. (2012). "A scalable method for solving high-dimensional continuous POMDP using local approximation." *arXiv preprint arXiv:1203.3477.*

Howard, R. A. (1960). *Dynamic programming and Markov Processes* Ronald A. Howard graph. Darst.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). "Planning and acting in partially observable stochastic domains." *Artificial intelligence*, 101(1-2), 99–134.

Kim, S. and Frangopol, D. M. (2010). "Cost-based optimum scheduling of inspection and monitoring for fatigue-sensitive structures under uncertainty." *Journal of Structural Engineering*, 137(11), 1319–1331.

Kochenderfer, M. J. (2015). *Decision making under uncertainty: theory and application*. MIT press.

Liu, W. (2017). "Optimizing flexible infrastructure designs via partially observable markov decision processes (POMDPs)." MSc Thesis, ERA, TU Munich.

Luque, J. and Straub, D. (2019). "Risk-based optimal inspection strategies for structural systems using dynamic bayesian networks." *Structural Safety*, 76, 68–80.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1987). "The complexity of markov decision processes." *Mathematics of operations research*, 12(3), 441–450.

Papakonstantinou, K. G., Andriotis, C. P., and Shinozuka, M. (2018). "POMDP and MOMDP solutions for structural life-cycle cost minimization under partial and mixed observability." *Structure and Infrastructure Engineering*, 14(7), 869–882.

Pooya, A. and Pakdaman, M. (2018). "Optimal control model for finite capacity continuous MRP with deteriorating items." *Journal of Intelligent Manufacturing*, 1–13.

Pozzi, M., Memarzadeh, M., and Klima, K. (2017). "Hidden-model processes for adaptive management under uncertain climate change." *Journal of Infrastructure Systems*, 23(4), 04017022.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Ross, S., Pineau, J., and Chaib-draa, B. (2008). "Theoretical analysis of heuristic search methods for online POMDP." *Advances in neural information processing systems*, 1233–1240.

Shani, G., Pineau, J., and Kaplow, R. (2013). "A survey of point-based POMDP solvers." *Autonomous Agents and Multi-Agent Systems*, 27(1), 1–51.

Straub, D. (2004). *Generic approaches to risk based inspection planning for steel structures*, Vol. 284. vdf Hochschulverlag AG.

Straub, D. and Špačková, O. (2016). "Optimizing adaptable systems for future uncertainty." *14th International Probabilistic Workshop*, Springer, 3–13.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). "Policy gradient methods for reinforcement learning with function approximation." *Advances in neural information processing systems*, 1057–1063.

Vahidi, A. (2018). "Stochastic methods for policy optimization in a decision problem: Investigating the efficiency of heuristics for solving partially observable markov decision processes (POMDPs)." MSc Thesis, ERA, TU Munich.

Špačková, O. and Straub, D. (2017). "Long-term adaption decisions via fully and partially observable markov decision processes." *Sustainable and Resilient Infrastructure*, 2(1), 37–58.